DOCUMENT RESUME

ED 446 595 HE 033 435

AUTHOR Anderson, M. Brownell, Ed.
TITLE Research in Medical Education: Proceedings of the Annual Conference (39th, Chicago, Illinois, October 30-November 1, 2000).
INSTITUTION Association of American Medical Colleges, Washington, DC.
ISSN ISSN-1040-2446
PUB DATE 2000-10-00
NOTE 155p.; Conference sponsored by the Association of American Medical Colleges in conjunction with its Annual Meeting (111th, Chicago, IL, October 27-November 1, 2000).
AVAILABLE FROM Publication Orders, Association of American Medical Colleges, 2450 N Street, N.W., Washington, DC 20037 ($25 plus $6 shipping). Tel: 202-828-0416. For full text: http://www.academicmedicine.org.
PUB TYPE Collected Works - Proceedings (021) -- Collected Works - Serials (022) -- Reports - Descriptive (141)
JOURNAL CIT Academic Medicine; v75 n10 suppl Oct 2000
EDRS PRICE MF01/PC07 Plus Postage.
DESCRIPTORS College Entrance Examinations; Higher Education; Licensing Examinations (Professions); *Medical Education; Medical Schools; Medical Students; *Program Evaluation; *Research; Trend Analysis

ABSTRACT

This collection contains research papers selected for presentation at a conference. Papers include the following: (1) "Morning Report: Focus and Methods over the Past Three Decades" (Z. Amin, J. Guajardo, W. Wisniewski, G. Bordage, A. Tekian, and L. G. Niederman); (2) "Context, Conflict, and Resolution: A New Conceptual Framework for Evaluating Professionalism" (S. Ginsburg, G. Regehr, R. Hatala, N. McNaughton, A. Frohna, B. Hodges, L. Lingard, and D. Stern); (3) "Tracking Knowledge Growth across an Integrated Nutrition Curriculum" (C. Hodgson); (4) "Following Medical School Graduates into Practice: Residency Directors' Assessments after the First Year of Residency" (G. L. Alexander, W. K. David, A. C. Yan, and J. C. Fantone, III); (5) "The Impact of an Alternative Approach to Computing Station Cut Scores in an OSCE" (J. H. McIlroy); (6) "An Investigation of the Impacts of Different Generalizability Study Designs on Estimates of Variance Components and Generalizability Coefficients" (L. A. Keller, K. M. Mazor, H. Swaminathan, and M. P. Pugnaire); (7) "A Validity Study of the Writing Sample Section of the Medical College Admission Test" (M. Hojat, J. B. Erdmann, J. J. Veloski, T. J. Nasca, C. A. Callahan, E. Julian, and J. Peck); (8) "Prediction of Students' Performances on Licensing Examinations Using Age, Race, Sex, Undergraduate GPAs, and MCAT Scores" (J. J. Veloski, C. A. Callahan, G. Xu, M. Hojat, and D. B. Nash); (9) "Does Institutional Selectivity Aid in the Prediction of Medical School Performance? (A. V. Blue, G. E. Gilbert, C. L. Elam, and W. T. Basco, Jr.); (10) "The Presence of Hospitalists in Medical Education" (J. A. Shea, J. S. Wasfi, K. J. Kovath, D. A. Asch, and L. M. Bellini); (11) "Dual Degree MD-MBA Students: A Look at the Future of Medical Leadership" (W. W. Sherrill); (12) "A Preliminary Analysis of Different Approaches to Preparing for the USMLE Step 1" (R. A. Thadani, D. B. Swanson, and R. M. Galbraith); (13) "Effectiveness of Telehealth for Teaching Specialized Hand-assessment

Techniques to Physical Therapists" (W. Barden, H. M. Clarke, N. L. Young, N. McKee, and G. Regehr); (14) "A Controlled Trial of an Interactive, Web-based Virtual Reality Program for Teaching Physical Diagnosis Skills to Medical Students" (J. A. Grundman, R. S. Wigton, and D. Nickol); (15) "Evaluation of a CME Problem-based Learning Internet Discussion" (J. M. Sargeant, R. A. Purdy, M. J. Allen, S. Nadkarni, L. Watton, and P. O'Brien); (16) "Correlates of Physicians' Endorsement of the Legalization of Physician-assisted Suicide" (K. D. Novielli, M. Hojat, T. J. Nasca, J. B. Erdmann, and J. J. Veloski); (17) "Learning Adolescent Psychosocial Interviewing Using Simulated Patients" (K. Blake, K. V. Mann, D. M. Kaufman, and M. Kappleman); (18) "Have Clinical Effectiveness Ratings Changed with the Medical College of Wisconsin's Entry into the Health Care Marketplace?" (D. Bragg, R. Treat, and D. E. Simpson); (19) "Six-year Documentation of the Association between Excellent Clinical Testing and Improved Students' Examination Performances" (C. H. Griffith, III, J. C. Georgesen, and J. F. Wilson); and (20) "When Residents Talk and Teachers Listen: A Communication Analysis" (J. L. Paukett). (SLD)

October 2000 • Volume 75 • Number Ten • Supplement

ACADEMIC MEDICINE

Research in Medical Education

Proceedings of the Thirty-ninth Annual Conference

Chicago, IL
October 30–November 1, 2000

Sponsored by the Association of American Medical Colleges
in conjunction with the 111th Annual Meeting
October 27–November 1, 2000

Research in Medical Education

Proceedings of the Thirty-ninth Annual Conference

Chicago, IL
October 30–November 1, 2000

Sponsored by the Association of American Medical Colleges
in conjunction with the 111th Annual Meeting
October 27–November 1, 2000

PUBLICATION OF THE ASSOCIATION OF AMERICAN MEDICAL COLLEGES
EDITORIAL OFFICE AND STAFF
Association of American Medical Colleges
2450 N Street, N.W., Washington, DC 20037; (202) 828-0590
EDITOR: Addeane S. Caelleigh
SENIOR DEPUTY EDITOR: Albert Bradford
DEPUTY EDITOR: Lisa Dittrich
STAFF EDITOR: Ann Steinecke
PEER-REVIEW MANAGER: Newton Holt
SENIOR EDITORIAL ASSISTANT: Amy Ciok
EDITORIAL ASSISTANT: Bridgett Ellison
PUBLISHING SERVICE: Hanley & Belfus, Inc.
PUBLISHER'S EDITOR: Sandra A. Lovegrove

www.academicmedicine.org

Special thanks to M. Brownell Anderson, associate vice president, Division of Medical Education, at the AAMC, for her excellent work as editor and coordinator of these Proceedings; and to the members of the 1999 and 2000 RIME committee for their collegial and cooperative work on this joint project.

•

Correspondence: Address all correspondence regarding the content of these Proceedings to M. Brownell Anderson, Associate Vice President, Division of Medical Education, Association of American Medical Colleges, 2450 N Street, N.W., Washington, DC 20037.

Permission to reproduce copies: For permission to reproduce this supplement for noncommercial scholarly use, address correspondence to the Office of the Editor, *Academic Medicine*, Association of American Medical Colleges, 2450 N Street, N.W., Washington, DC 20037.

To order: The price is $25 plus $6.00 for shipping. Send orders to Publication Orders, Association of American Medical Colleges, 2450 N Street, N.W., Washington, DC 20037. For further information about ordering, telephone 202-828-0416.

•

Academic Medicine (ISSN 1040-2446) is owned and published by the Association of American Medical Colleges. Publishing Services provided by Hanley & Belfus, Inc., 210 South 13th Street, Philadelphia, PA 19107. Second-class postage paid at Washington, DC, and additional mailing offices.

Published as a supplement to *Academic Medicine*, October 2000.

FOREWORD

Many of us eagerly look forward each year to receiving the October supplement to *Academic Medicine* (now available online at (http://www.academicmedicine.org)). We know that it will contain research reports and reviews that expand our knowledge and inspire us to encourage others to undertake the study of important questions as well as to continue our own research programs. The proceedings of the 39th annual conference on Research in Medicine Education (RIME) is no exception. The theme of this year's meeting, Making a Difference, is truly the goal of medical education research, and many facets of this goal are reflected in the papers that constitute the program for this meeting.

The purpose of the RIME conference is to provide a forum for the presentation and discussion of research concerning all aspects of medical education. The annual meeting of the Association of American Medical Colleges represents the largest gathering of the academic medicine community in the world. The meeting provides an opportunity for medical education researchers to demonstrate the vigor and diversity of scholarly investigations in medical education.

I am exceptionally pleased that Dr. Whitney Addington, director of the Rush Primary Care Institute and professor of Family Medicine, Internal Medicine, and Nursing, will give the 2000 RIME conference's invited address, entitled "Lifelong Learning: When Are We Going to Practice What We Preach?" Dr. Addington has promised to be provocative in his advocacy for continuous professional development. As the outgoing president of the American College of Physicians, he is in a unique position to reflect on this issue and challenge us to increase our inquiry on how best to accomplish the goals of lifelong learning. Dr. Addington is president of the Chicago Board of Health and has been a strong campaigner for universal health insurance in this country. I am delighted he has agreed to address us and I look forward to his comments.

An increased number of research papers were submitted to the RIME committee this year, another indication of the health of medical education research. Of the 98 research papers submitted, 36 were selected for presentation and inclusion in the proceedings. The topics range from the tried and true, such as standardized patient and OSCE examinations and related measurement issues, to some that have had limited discussion, such as telemedicine and Web-based education, the role of hospitalists, and the impact of the marketplace on academic medicine. Each paper is scheduled, along with one or two other papers on a similar topic, to a session led by a moderator, who has the opportunity to raise pertinent questions and ensure a wide-ranging discussion by the audience. In an effort to highlight topics that may be of special interest, six of the research papers were selected for presentation in two plenary sessions. This year, some of our outstanding researchers will be joined by a number of new faces. Especially gratifying is that several of these new faces are those of resident physicians and medical students, a further indication of the vitality of medical education research.

In addition to the research papers, there were 220 abstract proposals received, 96 of which were selected for presentation as either posters or oral presentations. As in the past, the RIME conference will host a reception to showcase the abstracts in the poster format on Monday evening. The oral abstracts will be presented in sessions organized around specific topics, each guided by a moderator.

This supplement to *Academic Medicine* includes, in addition to the research papers, two review papers and the invited address from 1999. Dr. Zubair Amin will review issues related to a common occurrence in residency education, the morning report, and point out areas that need to be the focus of research. The other review, presented by Dr. Shiphra Ginsburg, will address the challenge of evaluating professionalism and explore the relationships among context, conflict, and resolution in this endeavor. In 1999, Dr. Charles Friedman, from the University of Pittsburgh, presented a thought-provoking address on the role of informatics and information technology in medical education. His informative presentation, entitled "The Marvelous Medical Education Machine," is published in these proceedings.

This year, for the first time, the supplement also includes the 1999 Jack Maatsch Memorial Presentation, sponsored by the Office of Medical Education Research and Development at Michigan State University. "The Epistemology of Clinical Reasoning: Perspectives from Philosophy, Psychology, and Neuroscience," by Dr. Geoffrey R. Norman of McMaster University, draws on differing disciplines to expand the subject of clinical reasoning and its avenues for further research. It is accompanied by "Clinical Problem Solving and Decision Psychology," by Dr. Arthur Elstein of the University of Illinois at Chicago.

Two excellent symposia were selected for presentation, one, moderated by Dr. Deborah Simpson, on the challenging task of measuring faculty development outcomes, and the other, moderated by Dr. David Newble, on assessing the performances of practicing physicians. The always-popular "RIME wrap-up" session will round out the formal presentations. We are fortunate to have persuaded Dr. John Bligh, of The University of Liverpool, and editor of *Medical Education* (UK); Dr. Georges Bordage, University of Illinois at Chicago; and Dr. Judy Shea, University of Pennsylvania, to help us put the meeting's presentations in perspective and suggest where we might go from here

The RIME conference is planned and organized by the RIME committee, a committee of the AAMC's Group on Educational Affairs. On behalf of the committee, we wish to express our appreciation to all researchers who submitted papers for the meeting. It was not an easy task to select those to place on the program, and we are indebted to the essential contribution made by the external reviewers. These individuals provided suggestions and comments that have benefited the authors of the papers, symposia, and abstracts that make up this year's meeting.

The RIME committee wants to recognize the outstanding contribution of Brownie (M. Brownell) Anderson, Associate Vice-President, Division of Medical Education of the AAMC. I know the committee members join me in saying that the process of requesting and reviewing papers and developing the program was made infinitely easier by her guidance and wisdom, over and above her welcome sense of humor. Her staff, as well, eased our task. In addition, we appreciate the support of Addeane Caelleigh, editor of *Academic Medicine*, and her able staff for their assistance in publishing these proceedings. We hope you enjoy the meeting.

Beth Dawson
Chair, 2000 RIME Committee

RIME 2000 External Reviewers

We acknowledge with thanks the assistance of these reviewers for their constructive assessments of the papers submitted for the RIME conference.

Michael Ainsworth
Mark A. Albanese
Jerry H. Alexander
Sharon Allen
Elaine Alpert
Louise Arnold
Sara Axtell
Gwyn E. Barley
Barbara Barzansky
Bruce Bennard
Nancy Bennett
Eta S. Berner
James G. Boulger
Mary A. Bozynski
Carlos Brailovsky
Robert J. Bulik
Gwendie Camp
Sally H. Cavanaugh
Chris C. Cheesemen
Lynn Cleary
Steven G. Clyman
Jannette Collins
Jerry A. Colliver
Malcolm Cox
Louis L. Cregler
Joe E. Crick
Raymond H. Curry
Michael D. Cusimano
Joyce E. Dains
Debra A. DaRosa-Creek
W. Dale Dauphinée
Wayne K. Davis
Linda H. Distlehorst
Michael B. Donnelly
Janine C. Edwards
Mike Elnicki
Gary Stephen Ferenchick
Scott A. Fields
Andrew T. Filak Jr.
Rosemarie L. Fisher
James T. Fitzgerald
J. Roland Folse
Gail E. Furman
Erica Friedman
Alberto Galofre
Craig L. Gjerde
Gerald S. Gotterer
Evelyn C. Granieri
Jean D. Gray
Charles H. Griffith III
Louis Grosso
Paul L. Grover
Cyril M. Grum
Larry D. Gruppen
Hilary Haftel
Carol L. Hampton
Penelope A. Hansen
R. Van Harrison
Leo M. Harvill
Paul A. Hemmer
William D. Hendricson
Rebecca C. Henry
Susan Thompson Hingle
J. Dennis Hoban
Holly J. Humphrey
David M. Irby
Evelyn W. Jackson
Dr. Penny Jennett
Brian Jolly
Dorthea H. Juul
Summers Kalishman
David M. Kaufman
Daniel J. Klass
Regina A. Kovach
Sharon K. Krackov
Barbara M. Lawrence
Linda S. Lee
Lloyd A. Lewis
John H. Littlefield
John S. Lloyd
Nicole Lurie
Thomas Lynch
William D. Mattern
William C. McGaghie
Christine McGuire
James F. McKinsey
William P. Metheny
Bruce Z. Morgenstern
Gail Morrison
Hurley Myers
John Nolte
John J. Norcini
Geoffrey Norman
Ronald J. Nungester
Mark O'Connell
Curtis Olson
Patricia S. O'Sullivan
Rubens J. Pamies
Louis N. Pangaro
Tony Paolo
Steven Peitzmann
Linda C. Perkowskiool
Robert P. Perry
Henry S. Pohl
Kyle E. Rarey
Michael M. Ravitch
Glenn Regehr
Boyd F. Richards
Douglas Ripkey
Lynne S. Robins
John Rogers
Jonathan M. Rosen
Arthur I. Rothman
Robert F. Rubeck
M. Lynn Russell
Ajit K. Sachdeva
W. Scott Schroth
Thomas L. Schwenk
James L. Sebastian
John H. Shatzer
Judith A. Shea
Kent J. Sheets
Benjamin Siegel
Deborah E. Simpson
James M. Shumway
Stuart Slavin
Paula S. Smith
Jeannette E. South-Paul
Alex Stagnaro-Green
David J. Steele
Jeffrey A. Stearns
David T. Stern
David E. Steward
Steven A. Stiens
Barry Stimmel
Gregory N. Strayhorn
Christine A. Stroup-Benham
Jeffrey L. Susman
David B. Swanson
Robyn Tamblyn
Jeffrey Turnbull
Richard . Usatine
Steven J. Verhulst
Anthony E. Voytovich
Royce W. Waltrip II
David Warren
Kathleen Watson
Karen J. Wendelberger-Marcdank
Alison J. Whelan
Marcia Z. Wile
Brent C. Williams
Fredric M. Wolfe
Raymond Y. Wong
Moses K. A. Woode

6

2000
Research in Medical Education Conference Committee

Chair

Beth Dawson, PhD
Southern Illinois University School of Medicine

Past Chair

Ilene Harris, PhD
University of Minnesota Medical School–Minneapolis

Members

Louise Arnold, PhD
University of Missouri–Kansas City School of Medicine

Paul E. Mazmanian, PhD
Virginia Commonwealth University School of Medicine

Fredrick A. McCurdy, MD, PhD
University of Nebraska School of Medicine

S. Scott Obenshain, MD
University of New Mexico School of Medicine

Gordon Page, EdD
University of British Columbia Faculty of Medicine

Richard K. Reznick, MD
University of Toronto Faculty of Medicine

Sarah L. Stone, MD
University of Massachusetts Medical School

Executive Secretary

M. Brownell Anderson
Association of American Medical Colleges

7

Research in Medical Education
Proceedings of the Thirty-ninth Annual Conference
October 30–November 1, 2000

Chair: Beth Dawson, PhD

Editor: M. Brownell Anderson

Foreword by Beth Dawson, PhD

PAPERS

9

UC. 10

Author index appears on page S143

11

● REVIEW PAPER

Moderator: David Steward, MD

Morning Report: Focus and Methods over the Past Three Decades

ZUBAIR AMIN, JESUS GUAJARDO, WLODZIMIERZ WISNIEWSKI, GEORGES BORDAGE, ARA TEKIAN, and LEO G. NIEDERMAN

Residents rank morning report as the most important educational activity of their residency training.[1] Although there is a lack of documented evidence as to the educational value of morning report, the practice is ubiquitous across almost all primary care residency programs in North America. The ever-changing practice of medicine and ongoing demands for evidence in medical education force us to examine essential aspects of morning report in order to base future decisions about morning report on sound educational evidence. Thus, a systematic review of the published literature on morning report was done in order to identify the various purposes and modalities of morning reporr, to find evidence in support of its educational value, and to discuss possible future directions for research on morning report.

The term "morning report" is used to describe case-based conferences where residents, attending physicians, and others meet to present and discuss clinical cases. The term includes resident reports, morning or housestaff conferences, and morning sessions but excludes work rounds or teaching rounds. In a typical morning report, the team on duty during the night presents recently admitted patients, followed by a general discussion of the cases and related topics.

Data Collection

Data Identification and Study Selection. Four complementary approaches were used to locate articles about morning report. The goal was to retrieve all published articles. First, Medline, ERIC, and PsycINFO were searched using the key words morning report, morning session, residents' report, morning conference, education, and teaching. The key words wete used in various combinations and in different search modes (e.g., titles and subject headings). The search covered articles written between 1966 (start of Medline) and December 1999. No limitation was set on the search parameters. All journals, languages, and types of articles, including original articles, surveys, opinions, and letters to the editor, were included. Second, a manual search was conducted through non-indexed medical education journals. All relevant articles not previously identified by computerized searches were included. Third, the reference section of each article was reviewed and all pertinent articles not previously found were also retrieved and included for review. Finally, knowledgeable educators in the field were consulted in an effort to locate any additional articles not previously detected. As a result, 48 articles were found related to morning report. Although the search began with articles dating back to 1966, the oldest article on morning report was published in 1979. Most articles (80%) were published after 1990. Forty-one articles are discussed; seven other articles, mostly letters to the editor, addressed issues already covered elsewhere.[2-8]

Data Extraction. The selected articles were reviewed according to a three-step method as described by Gordon,[9] namely identification of key issues for review, selection of relevant information from various articles related to each issue, and critical synthesis and generalizations. The focus was primarily on the educational aspects of morning report and areas of possible improvement. We identified four major areas for review: purpose of morning report, organization, instructional methods, and educational outcomes. Each topic area is presented, followed by an overall discussion at the end.

Purpose of Morning Report

Historically, morning report probably was created to meet the demands of the hierarchical systems of public hospitals. In many cases, there were no ward attendings, and the chief of service had to ensure the health and safety of all the patients. Morning report provided the chief of service with the information needed to achieve this level of oversight.[10] Both the purpose and the audience of morning report have evolved over the years, and morning reporr is now conducted for diverse purposes with a wide variety of audiences. The various purposes were evident in the literature reviewed, with education becoming the main objective.[15] Other purposes were also mentioned, such as evaluating residents and the quality of services, detecting adverse events, and social interaction. The multiple purposes were evident in Parrino and Villanueva's sutvey of faculty and chief residents from 124 departments of medicine. Half of the respondents considered morning report "an important case-oriented teaching session" and a fifth believed that morning report "allow[ed] the chief of medicine or program director to keep tabs on medical services."[11] The importance of education was also reiterated in a recent survey where the majority of internal medicine residents indicated that education should be the primary purpose of morning report.[12] The various purposes of morning report are presented according to five subheadings: education, evaluation of residents and quality of services, detection and reporting of adverse events, non-medical issues, and social interaction.

Education. The educational goals pursued during morning report varied widely, ranging from case-based teaching[1,13-18] to reviewing and planning patient management,[1,15-17] fostering presentation skills,[15,19] highlighting the unique approach of the generalist physician,[19] developing intellectual curiosity and research,[15,19] promoting decision-making skills,[20] and self-directed learning.[20,21] Morning report was also used to teach residents selected topics that are not usually part of the curriculum, such as ethics.[22] Case-oriented teaching was the most frequently cited educational purpose of morning reports.[11]

Evaluation of Residents and Quality of Services. Most of the programs surveyed used morning report as a mean of evaluating residents' performances.[11-23] In Parrino and Villanueva's survey, faculty in many programs used morning report to evaluate residents' attitudes (84%), clinical skills (63%), and quality of care (93%).[11] A majority of respondents (82%) reported that morning report was also an effective means of case management.[11] Although morning report was used to evaluate residents and quality of care, no structured instrument or rating scale to conduct such evaluations was reported.

Detection and Reporting of Adverse Event. Morning report was sometimes used to detect and report adverse events.[24-26] Kaufmann reported that a pharmacy intern regularly attended morning report and considered whether admissions were related to medication

12

problems.[24] Sivaram et al. reported that adverse drug reactions were discussed in the business portion of morning report and were later reviewed by the Pharmacy and Therapeutic Committee.[25] Welsh et al. explored the effect of prompting residents to report adverse events.[26] All three studies concluded that morning report can be an effective means to detect and report adverse events such as drug reactions.

Non-medical issues. Although the discussion of non-medical issues during morning report was seldom reported, most programs addressed these issues on a regular basis. Schiffman et al. found that 85% of programs addressed a variety of non-medical issues such as social, personal, ethical, political, and economic topics, as well as cost–effectiveness and administrative matters.[27] Actual time spent on these issues during morning report was not reported.

Social Interaction. Although social interaction was not an explicitly stated goal, morning report provided an opportunity for residents and faculty to socialize. Eighty-five percent of the respondents in Parrino and Villanueva's survey indicated that morning report was an important social event for both residents and faculty.[11] Two thirds of the programs in Schiffmann's study served food and drinks during morning report and conducted business in an informal atmosphere that fostered social interaction.[27]

In summary, residency programs used morning report for multiple purposes, including education and a variety of other goals. Residents favor morning report as an educational activity. The relative importance of each purpose of morning report depends on individual programs and, in turn, may determine the way morning report will be organized and conducted.

Organization of Morning Report

Most of the articles that addressed the organizational aspects of morning report came from internal medicine residency programs. Other programs included pediatrics, family medicine, and neurology. The organization of morning report is presented according to five subheadings: frequency, time, and duration; participation, leadership, and tone; case selection and presentation; record keeping; and patient follow up.

Frequency, Time, and Duration. The frequency of morning report was fairly uniform across programs. Most were held on a regularly scheduled basis, with 80% of internal medicine programs holding morning report five times or more a week. Only a handful of programs held morning report less than three times a week.[27] Morning report usually began before 9 AM and lasted for an hour.[27] Some programs (4%) actually held "morning" report during the afternoon.[27] In most programs, work rounds preceded morning report to facilitate data collection prior to morning report. Schiffman et al. argued that conducting morning report after ward rounds may be more useful because attending physicians can contribute significantly to the quality of the session.[27]

Participants, Leadership, and Tone. The mix of participants and leaders varied greatly across programs. The chief of medicine or the director of medical education was present in more than half of the sessions.[27] Third-year service residents were the most regular participants, while the presence of first-year residents varied, with about 60% of the programs requiring their participation on a regular basis.[27] Gross et al. reported that internal medicine residents prefer the presence of generalist physicians at morning report, possibly because of the renewed interest in general internal medicine.[12] Carruthers described an Australian program where general practitioners from the community regularly attended morning report. She argued that a more widespread participation of general practitioners during morning report would lead to a better understanding of the strengths and weaknesses of general practice.[28] Finally, the presence of non-physician participants helped to broaden the scope of knowledge and experience of the residents. For example, pharmacists increased the detection of adverse drug reactions[24,25] and librarians increased the use of online searches by residents.[11] Some have argued against the presence of non-service personnel, junior residents, or medical students at morning report because their presence might inhibit the spontaneity of case presentation and discussion.[27]

Studies of verbal interactions during morning report consistently showed that participants tend to be rigid in their roles and in their ways of asking for or providing information. Most of the information exchanged was low-level factual information. Few questions were asked that required synthesis of patient information and medical knowledge.[29,30]

The person leading morning report was either a faculty member (70%) or a chief resident (30%).[11] Many openly criticized the role of the leaders and the tone they set during morning report.[10,19,31,32] Comments such as "morning retort or morning distort," "where bottom line is style above substance,"[31] and "secretive closed-door session"[12] were reported frequently. McGaghie et al. described the menacing atmosphere that prevailed in one institution as "... housestaff defining and defending mishaps using mechanisms such as denials, discounting, and distancing."[32]

Case Selection and Presentation. The selection and mode of presentation of cases also varied greatly among programs, reflecting most often the chief resident's or attending physician's preferences.[27] Case presentations varied from brief presentations of all cases with equal emphasis on each case to elaborate presentations of one or two "interesting" cases. Accordingly, times allotted for each case presentation varied widely. Westman prospectively compared the nature of the cases presented in internal medicine at a university center with those at an affiliated Veterans Administration hospital. The case mixes were similar in the two institutions; most cases (88%) were those of inpatients.[33] Gerard et al. reported that pediatrics residents were more likely to select cases whose diagnosis changed during hospitalization.[34] Other unorthodox methods of case selection and presentation included the selection of cases one to two days in advance,[35] the selection of simple cases at the beginning of the academic year and more complex ones later in the year,[27] and the presentation of cases prior to discharge.[20,36]

Record Keeping. Record keeping was done for different purposes during morning report.[15,17,18,27,37,38] Records were kept for educational purposes, such as the evaluation of content coverage[33] and patient follow ups,[15] or as data sources for research.[17] The availability of computers enabled many programs to use the data from morning report for a variety of purposes. Rouan et al. described a computer program to generate information from hospital admissions. They used the information for patient follow up, patient distribution among housestaff, residents' evaluation, and quality assurance.[37] Recht et al. also described a computerized data management program and its use in clinical research and quality assurance.[17]

Patient Follow Up. Most internal medicine programs allowed for patient follow ups.[27] Wegner and Shpiner showed that a final diagnosis was not always available at the time of discharge.[18] Similarly, Barton et al. compared pediatrics morning reports from a community hospital and a university hospital. In both settings, significant numbers of patients, 28% and 58%, respectively, were not diagnosed at the time of presentation at morning report.[19] Both investigators concluded that provision of patient follow up in morning report was important to maximize education.

In summary, there was a fair amount of regularity and similarity among programs in the frequency, time, and duration of morning report. There was more variability in the mix of participants and leaders, case selection, record keeping, and patient follow up. Many openly criticized the type of leadership used in conducting morning report. There was a lack of evidence in the literature on how the different purposes of morning report might affect its organization and the educational and clinical outcomes.

Instructional Methods

The most frequent instructional method used during morning report was case-based presentation, followed by discussion. Over three fourths of the programs surveyed by Malone and Jackson used such an approach.[40] Variations of case-based presentations were also used in an effort to improve educational effectiveness. For example, the chairman and chief resident would meet prior to morning report to review cases and preselect critical points for discussion.[15] The limitations of case-based presentations were also discussed in the literature, most notably by Parrino and Villanueva,[11] Mehler et al.,[41] and Hill et al.[42] Mehler et al. argued that "the standard format of case presentation may be less than optimal and can become a hackneyed experience."[41] Some shortcomings of case-based presentations have been addressed through innovative methods such as the presentation of prepared topics, photographic materials,[43] and learner-centered learning approaches.[40] In learner-centered approaches, the residents would determine the goals of the session once the cases were presented and then formulate questions based on these goals.[40] Parrino and Villanueva further proposed that "new techniques at morning report could be based on existing models of problem-based learning."[11] Battinelli echoed this view and advised learners to be creative and try new approaches.[44]

Like medical education, morning report faces a dilemma over its educational focus. Two main orientations emerged from the review. One focused on the need to increase the residents' knowledge level, the other on the need to improve their problem-solving and data-gathering skills. DeGroot and Siegler described the dilemma by using the analogy of the retentive "sponge mode" versus the inquisitive "search mode."[19] Years later, Richardson and Smith revisited this issue and reemphasized the importance of learning the process of information gathering and analysis rather than simply acquiring content knowledge.[45] Reilly and Lemon described a four-phase (similar to evidence-based medicine) morning report to foster active learning.[46] The first phase was devoted to the discussion of assigned questions from the previous day. Next, residents briefly presented all admission cases and the chief resident used didactic methods to emphasize important teaching issues. The participants then discussed in detail one particular case chosen for its educational value. Finally, the last five minutes were spent on formulating questions and assigning them to residents for presentation the next day. Reilly and Lemon reported a department-wide, positive impact following the introduction of this format. In addition, residents learned the principles and procedures of evidence-based medicine and how to formulate precise and clinically relevant questions.

Educational Outcomes

In an era of evidence-based medicine, evidence is also needed in education to enlighten existing educational practices and to plan new ones. Half of the 48 articles on morning report (52%) were based on studies. Surveys and questionnaires were used most often to collect data (nine studies); other data-gathering methods were observations, video recordings, quizzes, logbooks, and hospital records. Most studies were based on single programs; only four were conducted with multiple programs.[11,12,17,27] Some articles were based on anecdotal reports without any detailed data presented.

Wartman stated that detailed discussions, chart reviews, and analysis of hospital bills of selected discharged patients resulted in subsequent reductions in lengths of stay and controllable costs.[20,36] Similarly, Mehler et al. described a model of morning report that resulted in less test ordering and fewer requests for consults.[41] They reported that the participants' level of enthusiasm declined during the academic year and that more in-depth discussion of single cases became more attractive as time went on. Bassiri et al. introduced changes in morning report—such as presentation of articles, comments by specialists, a computer database, and regular followups—that improved the level of discussion and generated data for research.[14] Potyk et al. reported that both quizzes and mini-lectures increased learning, as measured by a true–false test administered later, although the quiz format resulted in better information retention.[47] D'Allessandro and D'Allessandro reported the use of radiology slides at pediatrics morning report as a means of increasing residents' interest.[48] Finally, several authors reported that morning report covered a broad range of topics included in published curricula (e.g., Pediatrics Review and Education Program by the American Academy of Pediatrics)[49] and in major medical references (e.g., internal medicine textbooks).[16] All programs that implemented innovations reported positive results as measured by increases in residents' knowledge[47] or desired behaviors.[15,24-26]

Discussion

Some key findings emerged from the diverse, albeit limited number of, publications on morning report (48 articles over 20 years). First, the purposes of morning report varied widely, although education was most frequently cited and favored by residents. Other important purposes were also mentioned, such as patient management and program and resident evaluation. Second, certain characteristics of the organization of morning report, such as frequency, timing, and duration, were fairly similar across programs. On the other hand, mix of participants, case selection and presentation, leadership, record keeping, and patient followup varied widely across programs. Tone, leadership, and the learning environment were often criticized. Third, various interventions that were implemented to improve the educational and clinical outcomes of morning report generally resulted in positive and promising results, although further validation of these findings is needed. Fourth, most of the published studies were from single programs, especially in internal medicine. There were very few studies on medical students and morning report. Encouragingly, there is renewed interest in morning report as an educational activity, as evidenced by the steady growth of published articles during the past decade.

The limited evidence available on morning report makes it difficult to make grounded recommendations, but some of the models used to plan and implement morning report were based on sound educational principles. For example, Reilly and Lemon's model of morning report is unique in that it encourages active learning, maintains continuity, and improves research activities in the program.[46] Such theory-based models can serve as the foundation on which to develop sound educational interventions that can be submitted to the scrutiny of the educational researchers. There is a clear lack of studies to document the effectiveness of morning report. This paucity may be due to the difficulties of doing research in the context of a multifaceted and multifactorial situation such as the multiple purposes, organizations, and audiences involved in morning report. It is also difficult to isolate the effects of morning report from those of other formal and informal educational activities. Finally, the lack of validated assessment instruments also adds to the difficulty of doing research on morning report. These difficulties should not be seen as insurmountable obstacles but as challenges to be met.

Future research is needed in four key areas. First, there is a need to characterize the types of learning and teaching that go on during morning report. What are the unique teaching and learning characteristics of morning report compared with other educational activities such as work rounds or teaching rounds? Second, little is known about the satisfaction levels of participants and the motivational factors that are operative during morning report. Although residents value morning report as their most important day-to-day learning activity, they also harbor strong negative feelings about the atmosphere that prevails. Could the quality of morning report be enhanced by analyzing more closely the positive and negative

feelings of the residents and the faculty? Research is also needed to document the effects of morning report on residents' knowledge, behaviors, and attitudes, as well as on patients' health care outcomes. Finally, there is a need for multi-institutional research on the effectiveness of new strategies to conduct morning report in order to verify the robustness of the interventions and thus move beyond program-specific effects.

Although the main focus of morning report has been on inpatient topics, there is a need to address the specifics of morning report in the context of ambulatory care. The pioneering work by Malone and Jackson indicated that the educational characteristics of ambulatory morning reports are significantly different from those of inpatient morning reports.[40] Consequently, simple generalization of results from inpatient modalities to ambulatory care is not recommended. Ambulatory morning report is relatively new and offers ample opportunities for high-quality research, including the identification of the specific learning needs of the participants. What are the unique components of the residents' education that should and can be addressed during ambulatory morning report? What are the unique educational attributes of ambulatory morning report? How can the continuity between ambulatory morning report and inpatient morning report best be ensured? Other priority research areas include studies of the natures of the cases presented and their relationships to educational and clinical outcomes.

The majority of studies on morning report came from internal medicine programs, with only a handful of reports from pediatrics, family medicine, and surgery. There is a need to plan studies across specialties to inform one another about the effectiveness of the innovations. Although morning report is primarily focused on residents, there are other important participants present during morning report, such as medical students, ethicists, and pharmacists. There was little focus in the literature on the participation of these types of participants during morning report. The educational needs and learning characteristics of this diverse audience are different from those of residents and need to be studied as well.

Morning report is a time-honored tradition. It is not just a ritual of early morning social gathering or a one-stop opportunity for program directors to keep tabs on the program. It is a valued time for residents, an uninterrupted flow of priceless minutes set aside from the hectic morning schedule for learning. Morning report is an opportunity for residents to exercise and improve their knowledge and their leadership, presentation, and problem-solving skills. Yet reports of its educational effectiveness are mostly anecdotal and its purpose often implicit or not explicitly defined. Each individual program must decide what it wants to achieve with morning report and structure the activity accordingly, distinguishing it from sitting rounds or patient-management rounds. Research is needed to document the educational and clinical effectiveness of morning report and to assess the relative merits of various ways of conducting morning report such that evidence and tradition can go hand in hand.

This review was done as part of an Independent Study while Drs. Amin, Guajardo, and Wisniewski completed a masters' degrees in health professions education in the Department of Medical Education at the University of Illinois at Chicago and were fellows in the International Educational Partnership in Pediatrics program jointly administered by the Department of Pediatrics and the Department of Medical Education.

Correspondence and requests: Zubair Amin, MD, MHPE, K. K. Women's and Children's Hospital, 100 Bukit Timah Road, Singapore 229899; e-mail: ⟨zubair@kkh.com.sg⟩.

References

1. Ways M, Kroenke K, Umali J, Buchwald D. Morning report. A survey of resident attitudes. Arch Intern Med 1995;155:1433–7.
2. Weitberg AB. The morning-report syndrome. N Engl J Med. 1980;302:925.
3. Adams K, Bennett M, Gooch B. The morning-report syndrome. N Engl J Med. 1980;302:925.
4. Bronson DL, Bertsch TF. Morning report for internal medicine clerks. Acad Med. 1993;68:780.
5. Nair BR, Hensley MJ, Pickles RW, Fowler J. Morning report: essential part of training and patient care in internal medicine. Aust NZ J Med. 1995;25:740.
6. Parrino TA. Recapturing immediacy in morning report. Ann Intern Med. 1994; 120:442–3.
7. Soubani AO. Morning report: a chief resident's perspective. J Gen Intern Med. 1994;9:237–8.
8. Stockman JA 3rd. The morning report. Clin Pediatr. 1997;36:589–90.
9. Gordon MJ. Organizing and Managing an Interactive Review of the Literature. Seattle, WA: Department of Family Medicine, School of Medicine, University of Washington, 1993.
10. Parrino TA. The social transformation of medical morning report. J Gen Intern Med. 1997;12:332–3.
11. Parrino TA, Villanueva AG. The principles and practice of morning report JAMA. 1986;256:730–3.
12. Gross CP, Donnelly GB, Reisman AB, Sepkowitz KA, Callahan MA. Resident expectation of morning report: a multi-institutional study. Arch Intern Med. 1999, Sep;159:1910–4.
13. Barbour GL, Young MN. Morning report: role of the clinical librarian. JAMA. 1986;255:1921–2.
14. Bassiri A, Kassen BO, Mancini GB. Improving the format of morning report. Acad Med. 1995;70:342–3.
15. Pupa LE Jr, Carpenter JL. Morning report: a successful format. Arch Intern Med. 1985;145:897–9.
16. Ramratnam B, Kelly G, Mega A, Tilkemeier P, Schiffman FJ. Determinants of case selection at morning report. J Gen Intern Med. 1997;12:263–6.
17. Recht L, Kramer P, Schwartz W. Morning report in computer era: tradition meets technology. Med Teach. 1995;17:327–31.
18. Wenger NS, Shpiner RB. An analysis of morning report: implications for internal medicine education. Ann Intern Med. 1993;119:395–9.
19. DeGroot LJ, Siegler M. The morning-report syndrome and medical search. N Engl J Med. 1979;301:1285–7.
20. Wartman SA. Morning report revisited: a new model reflecting medical practice of the 1990s. J Gen Intern Med. 1995;10:271–2
21. Harris ED Jr. Morning report. Ann Intern Med. 1993;119:430–1.
22. Silverman HJ. Description of an ethics curriculum for a medicine residency program. West J Med. 1999;170:228–31.
23. Blank L, Grosso L, Benson JA Jr. A survey of clinical skills evaluation of practices in internal medicine residency programs. J Med Educ. 1984;59:401–16.
24. Kaufman MB. Drug reactions identified at resident physicians' morning report. Am J Health-System Pharm. 1995;52:2031.
25. Sivaram CA, Johnson S, Tirmizi SN, Robertson V, Garcia D, Sorrells E. Morning report: a forum for reporting adverse drug reactions. Joint Comm J Qual Improvement. 1996;22:259–63.
26. Welsh CH, Pedot R, Anderson RJ. Use of morning report to enhance adverse event detection. J Gen Intern Med. 1996;11:454–60.
27. Schiffman F, Mayo-Smith M, Burton M. Resident report: A conference with many uses. Rhode Island Med J. 1990;73:95–102.
28. Carruthers A. General practitioner participation in 'morning report' at a major teaching hospital. Aust Fam Physician. 1997;26:S96–8.
29. Foley R, Smilansky J, Yonke A. Teacher–student interaction in a medical clerkship. J Med Educ. 1979;54:622–6.
30. Osheroff JA, Forsythe DE, Buchanan BG, Bankowitz RA, Blumenfeld BH, Miller RA. Physicians' information needs: analysis of questions posed during clinical teaching. Ann Intern Med. 1991;114:576–81.
31. Brancati F. Morning distort. JAMA. 1991;266:1627.
32. McGaghie WC, Engel JD, Wolf K, Smith AC. Morning report: a descriptive view from two different academic settings. Proc Annu Conf Res Med Educ. 1985;24: 157–62.
33. Westman EC. Factors influencing morning report case presentations. South Med J. 1999;92:775–7.
34. Gerard JM, Friedman AD, Barry RC, Carney MJ, Barton LL. An analysis of morning report at a pediatric hospital. Clin Pediatr 1997;36:565–8.
35. Ahsan AM. Morning report. not just a matter of attitude. Arch Intern Med 1996; 156:685.
36. Wartman SA. A new morning report model. Acad Med. 1994;69:820.
37. Rouan G, Marks E, Tuchfarber B, Redington TJ, Lorenzi N. Development, effects, and educational outcomes of reporting from a clinical database. Med Decis Making. 1991;11:S80–9.
38. Schiffman FJ. Morning report and work rounds: opportunities for teaching and learning. Trans Am Clin Climatological Assoc. 1995;107:275–86.
39. Barton LL, Rice SA, Wells SJ, Friedman AD. Pediatric morning report. an appraisal. Clin Pediatr. 1997;36:581–3.
40. Malone ML, Jackson TC. Educational characteristics of ambulatory morning re-

15

port. J Gen Intern Med. 1993;8:512–4.
41. Mehler PS, Kaehny WD, Fraser V, et al. Clinical effect of morning report. Acad Med. 1993;68:547.
42. Hill RF, Tyson EP, Riley HD Jr. The culture of morning report: ethnography of a clinical teaching conference. South Med J. 1997;90:594–600.
43. Biébel M, Hill W, Rucka J. Case snapshots: a new tool for morning report. Acad Med. 1998;73:594.
44. Battinelli D. Morning Report. In: Chief Resident's Manual. Washington, DC: Association of Program Directors in Internal Medicine, 1995, 15–20.
45. Richardson W, Smith L. Turning the Table: Learning by Cooperative Reasoning in Morning Report. Raleigh, NC: Association of Program Directors in Internal Medicine Meeting, 1993.
46. Reilly B, Lemon M. Evidence-based morning report: a popular new format in a large teaching hospital. Am J Med. 1997;103:419–26.
47. Potyk D, Novan G, Palpant S, Auricchio RJ, Benson J, Watson P. Comparing two formats for clinical "pearls" at morning report. Acad Med. 1997;72:73–4.
48. D'Alessandro DM, D'Alessandro MP. Radiologic education of pediatric residents during morning report. Acad Radiol. 1997;4:534–8.
49. D'Alessandro DM. Documenting the educational content of morning report. Arch Pediatr Adolesc Med. 1997;151:1151–6.

16

● **REVIEW PAPER** *Moderator: Norma Wagoner, PhD*

Context, Conflict, and Resolution: A New Conceptual Framework for Evaluating Professionalism

SHIPHRA GINSBURG, GLENN REGEHR, ROSE HATALA, NANCY MCNAUGHTON, ALICE FROHNA, BRIAN HODGES, LORELEI LINGARD, and DAVID STERN

During medical school, students are taught the knowledge, skills, and attitudes required to become competent physicians. Knowledge and skills are rigorously evaluated by written and oral exams, standardized patient scenarios, and ward evaluations. However, evaluation of behaviors, including professionalism, is often implicit, unsystematic and, therefore, inadequate. This is problematic for several reasons. First, medical schools are doing a disservice to future postgraduate training programs, as well as to society, by not explicitly and accurately evaluating this area during medical school. It is recognized that more complaints against physicians to medical societies relate to unprofessional conduct than to lack of knowledge or poor technical skills.[1] Yet students who display unprofessional behavior may not be identified in the current system, and will be promoted academically on the basis of adequate performance on tests of knowledge and skills alone.[2,3]

Second, we are doing a disservice to our students by not providing explicit feedback in this domain, thereby missing valuable opportunities to bring about awareness and improvement. The American Board of Internal Medicine, in its report "Project Professionalism," discussed the problem of erosion of professionalism during medical training. While knowledge and skills improve markedly over the four years of medical school, there is ample anecdotal evidence, and substantial quantitative evidence, that professional behaviors can diminish over this period.[4–6] There appears to be an unrealistic expectation that students will arrive at medical school lacking in knowledge and skills, but with a full complement of appropriate behaviors that require no further attention. However, all students are vulnerable to lapses in professional behavior and can benefit from explicit, systematic attention in this domain. The focus of medical education in the past century was on knowledge and skills. For the future of medicine, attention to the teaching and evaluation of professionalism is vital.

While this need to evaluate professionalism effectively has been recognized for some time, traditional methods of addressing the problem have not been particularly successful, for several reasons. The traditional approach to this issue has involved the identification and definition of the attitudes and concepts that comprise the concept of professionalism (such as altruism, accountability, excellence, duty, honor, integrity, and respect). Evaluation methods that rely on such abstract and idealized definitions lead us to discuss *people*, rather than their *behaviors*, as being honest or dishonest, professional or unprofessional. This implies that professionalism represents a set of stable traits.

Interestingly, a large literature exists that suggests the opposite. Many studies in personality psychology have shown that the presence of specific personality traits does not predict behavior.[7,8] For example, in one study of psychiatry residents, Minnesota Multiphasic Personality Inventory testing revealed serious personality disorders in the two individuals who eventually lost their licenses for professional misconduct.[7] However, several other participants showed the same personality traits, yet had no difficulty reported in 15 years of follow up. Thus, evidence suggests that the identification of specific traits does not allow us to predict an individual's behavior.

There are several reasons why this issue is important when discussing the evaluation of professionalism. Stable trait measures do not take into account a recognition that behaviors enacted often involves an effort at resolving a conflict between two (or more) equally worthy professional or personal values. For example, it is easy to say that one must always tell the truth, and that one must always protect patient confidentiality. However, these values may occasionally come into conflict, and the ultimate choice the student makes will depend on the specifics of the situation.[9,10]

In addition, professional behaviors are known to be highly context-dependent.[10,11] One can imagine a basically honest person lying to a patient given a particular context. This does not automatically mean that that person is dishonest, and therefore unprofessional. Certainly in social situations, a decision to always tell the full truth would be considered highly inappropriate.

Although the issues of conflict and context are separate at a theoretical level, in day-to-day practice they are likely to interact. One study has shown that 87% of physicians surveyed indicated that deception is acceptable on rare occasions, for example, if the patient would be harmed by knowing the truth, in order to circumvent "ridiculous rules," or to protect confidentiality.[12] Yet, when two specific professional values are in conflict, it is not always predictable which of the two values will take precedence. For example, while it is sometimes appropriate to lie in order to protect patient confidentiality, there are circumstances in which it would be considered more appropriate to break confidentiality rather than tell a lie. As one participant stated, honesty is "usually" the best policy, but everything is taken on a case-by-case basis, and any actions taken depend on the specifics of the people and the situation.[12] Traditional ways of evaluating professionalism do not make allowances for these gray areas.

Another element of evaluating professionalism involves the process of resolving the conflict. The ultimate choice an individual makes, manifested as the behavior witnessed, does not tell us how he or she arrived at the decision. We know nothing of whether the student recognized the professional "values" that were in conflict, or why the student chose to act in that particular way. So while focusing on behaviors rather than personality or character traits is important, we must also attempt to understand the process that led to the behavior.

Thus, if we do not include conflict, context, and the process of resolution in our evaluation methods, we might not be able to conduct the most reliable, valid, and appropriate evaluation of these behaviors.

Another reason for the lack of success of traditional approaches is that evaluators have not been willing to identify an individual as unprofessional for actions that appear to be relatively minor. Thus, lapses in professional behavior tend to be ignored or suppressed, due to an understandable reluctance to apply the broad, harsh label of "unprofessional."[11] In one study, clinician supervisors admitted and demonstrated their reluctance to give negative feedback regarding unprofessional behavior, even though in interviews they had stated strongly that they would do so.[14] Even if faculty have this willingness, they have been found to have "difficulty in identifying problems, an inability to verify problems, and fear of litigation" that inhibit their reporting of behavioral problems.[2]

This outcome arises, in part, from the fact that educators and researchers have traditionally focused on this problem from an abstract perspective. The definitions and subcategories of the broader concept of professionalism describe the idealized person, the "con-

17

summate professional," with no room for mistakes. With this theoretical basis, if someone tells a lie, even for a "good" reason, he or she could be suddenly labeled "dishonest," and therefore, "unprofessional." The only thing left for the evaluator to decide, then, is how unprofessional the individual is. This top-down focus on professionalism as an abstraction rather than a bottom-up focus on professionalism as a set of actions in context, therefore, is flawed.

This paper elaborates on the issues around this problem. First, we review the literature on the types of evaluation instruments used for measuring professionalism in medical education. We then outline fundamental conceptual deficiencies that exist in this literature. We argue that the three most important missing components are: consideration of the contexts in which unprofessional behaviors occur, the conflicts that lead to these lapses, and the reasons students make the choices they make. We then propose strategies for resolving these issues.

Method

We conducted searches through Medline, Psychlit, and ERIC for literature published over the past 20 years. We included studies that contained original research on the topic of assessment or evaluation of professionalism in medical education, or included instruments to measure professional behavior, professionalism, humanism, behaviors, values, and attitudes. After initial articles were identified, bibliographies were used to identify additional references, and experts in the field were consulted for missing but relevant papers. This process uncovered few studies addressing specific efforts to evaluate professionalism. There was an abundance of articles calling for new and better methods of evaluation, and arguments for why this is so important and neglected. Some papers dealt with certain aspects of professionalism, for example, ethics, communication skills, interpersonal skills, and humanistic behavior, but they did so without extrapolation to the larger notion of professionalism. These studies were included if they highlighted difficulties in evaluating professionalism or provided new insights or solutions, and contained original research.

Results

Evaluations by Faculty Supervisors. In 1979, the AAMC interviewed approximately 500 clerkship directors about "problem students." They identified 21 types of problem students, and then asked how often each type of problem was seen, and how difficult the problem was. Among the results from the University of Washington School of Medicine, researchers found that "noncognitive" issues (e.g., bright but poor interpersonal skills) were "frequent and difficult," but that the very disturbing ones (e.g., cannot be trusted, manipulative) were seen only infrequently.[15] Though this survey was done many years ago, it provides an early glimpse of faculty's concerns about the professional behaviors of students. Since then, various other studies have analyzed approaches used by faculty in the evaluation of professionalism, including global rating scales, in-training evaluations, and encounter cards.

Ward rating forms, completed by the physician–supervisor, are the most commonly used instruments. In addition to assessing medical knowledge and clinical skills, many of these forms have a single global item to assess professional behavior, which may be subject to extensive rater bias.[16,17] A study by Woolliscroft et al. highlights some of the problems of using this type of assessment. The authors found that using a questionnaire, faculty could assess the humanistic qualities of internal medicine residents, at least for the item "doctor–patient relationships."[18] However, it would take 20–50 faculty members per resident to achieve acceptable reproducibility, which calls into question the utility of this instrument. This also suggests that the trait doctor–patient relationships is probably not stable, but rather may be subject to context bias. Different evaluators might see different behaviors or make different interpretations. In a related study, Johnson found that physicians' and nurses' evaluations of intensive care unit residents correlated highly with respect to all criteria *except* the assessment of humanistic qualities, further highlighting the importance of context.[19]

To compensate for the problem of infrequent observations, systems have been developed that encourage the repeated observation and documentation of the performances of medical trainees (often on a daily or weekly basis).[20,21] This allows for the assessment of knowledge, skills, or professional behaviors with reasonable inter-rater reliability and construct validity. Such real-time evaluations permit early intervention, facilitate feedback, and guide remediation. However, in a study of encounter cards in the evaluation of anesthesia residents, despite numerous negative comments by supervisors, only 1% of the comments were found to be about unprofessional behaviors.[22] Further, those residents who received these negative comments were only rarely rated overall as "performing below level" by their supervisors, despite their all having had critical incident reports and scoring lower on objective testing. This, again, highlights the difficulties faculty have in documenting unprofessional behavior.

Faculty can, in fact, be trained to accurately observe and assess specific behaviors. One group developed a reliable assessment of a very specific set of humanistic skills (e.g., introduced self to the patient, acknowledged the agenda from the last visit) by asking faculty to view videotapes of residents' interactions with patients.[23] However, even if faculty can identify problematic behavior in a reliable way, they are often reluctant to record it. Burack, using a rigorous qualitative method, demonstrated that faculty have a marked reluctance to respond unambiguously to behaviors that indicate negative attitudes towards patients.[14] In interviews, faculty stated that they would not tolerate "this sort of behavior" and would "definitely lay down the law" if such behavior were observed. However, in practice they usually did not respond at all, or did so in such a way as to require interpretation by the learner. The feedback can then be misinterpreted to be permissive. As explanations for this dichotomy, clinicians reported their sympathy for the learners' stress, as well as the possible penalties educators can face for giving negative feedback, such as receiving bad teaching evaluations and being open to personal and legal risks. They felt that if the observed behavior is only a lapse, and the learner is fundamentally "good," corrective feedback might discourage or frustrate the resident. Conversely, for fundamentally "bad" residents, corrective feedback is seen as futile.

Therefore, methods that exist for faculty evaluation of professional behavior are problematic. Evaluations cannot be kept on theoretical, abstract, or definitional levels; thus, these scales have poor reliability. Numerous observations in various contexts need to be made, but attending physicians are present for only a small proportion of the time. In addition, even when lapses in professional behavior are identified, there is great reluctance to report them.[14]

Nurses and Patients. Some of the reluctance faculty have in evaluating professional behavior results from potential conflict in their roles as teacher, mentor, and evaluator. Other groups, such as patients[24,25] or nurses,[19,26,27] may not be subject to these conflicts. In addition, these other groups may see the students and residents more often and in different contexts. Woolliscroft's study included groups of nurses and patients; unfortunately, the patients' ratings were not reliable, and it would have required up to 50 patients' assessments to achieve a reproducible estimate of professional behavior.[18] Nurses achieved good reproducibility with ten to 20 assessments per resident, but this amount may still be impractical. Because professional behavior is so context-specific, it is not surprising that only low to modest correlations exist between ratings by these different assessors. Also, nurses and patients may face different kinds of pressures that could deter their unbiased reporting of unprofessional behaviors; for example, a patient may be reluctant to jeopardize the continuity of a relationship with a physician even though it is problematic. In addition to highlighting some of the

18

difficulties in evaluating professional behavior, Woolliscroft et al.'s study provides a good example of an attempt to triangulate results as a measure of validity.

Peer Evaluation. Peers are in a good position to evaluate each other's professional behaviors because of frequent, close, and varied contact. Thus, the use of peer assessment of professional behaviors may solve many of the problems described for faculty's assessment. However, several problems remain and some new problems may arise through the use of peer assessment.

On a positive note, there is some suggestion that medical students' peer evaluations may be the best measures of interpersonal skills available.[28–30] Thomas et al. reported a pilot study of peer review in residency training using a ten-item questionnaire.[31] The items on the form clustered into two domains: "technical skills" and "interpersonal skills," which included humanistic behaviors. Of particular interest is this study's finding that intern peer evaluations of a composite "professionalism" domain correlated well with faculty evaluations of the same dimension ($r = .57$, $p < .05$). An interesting modification of a ranking system that forces students to discriminate among their peers based on certain dimensions of professionalism has been described.[32] The authors suggest that such a system enables identification of the top 10–15% of the class, but it is not helpful in discriminating among the rest, perhaps because the students were asked for only positive nominations on the peer-evaluation form.

On the other hand, peers, like faculty, seem to have a difficult time discriminating the abstract dimensions of professionalism from each other and from other skills. For example, in a study of peer assessment of professional dimensions, Arnold found very high internal consistency (coefficient alpha) across the dimensions, suggesting a strong halo effect in the ratings of the separate dimensions.[29] Further, scores were highly correlated with more knowledge-based measures such as National Board of Medical Examiner's exam (Parts I and II) and grade-point average, suggesting that dimensions other than professionalism were also contributing to the scores. Also, as with faculty ratings, it would appear that a fairly large number of ratings are necessary to obtain stable measures across raters.[33,34] Interestingly, the numbers of negative peer evaluations generated in the small groups depended upon the kind of faculty leadership exercised in each group.[29] This constitutes yet another example of the importance of context and social climate in peer (and other) assessment methods.

In fact, the social climate of peers assessing peers may have negative consequences. That is, while some studies report positive reception of peer feedback, others report marked resistance to peer evaluation even though the evaluations were anonymous and for research purposes only.[31,35] Helfer found that senior medical students were more accepting of peer evaluations than were junior students, who lacked confidence in the usefulness of the system.[30] Van Rosendaal found that residents worried that the process would undermine their work and personal interrelationships.[35]

In summary, peer evaluations hold promise for evaluating professionalism. However, before they are likely to be very useful, many of the same problems facing faculty's evaluation of professionalism will have to be solved, and evaluation systems must be developed that will overcome the reluctance of peers to rate one another.

Self Evaluation. Several early studies were conducted that involved self-reports of attitude changes during medical training. To varying degrees, these students reported increases in certain attitudes, such as cynicism; were more concerned about making money; or felt that their ethical principles had become eroded or lost.[5,6,36,37] Some positive attitudes increased as well, for example, concern for patients, and helpfulness.[5] More recently, Clack studied gender differences in medical graduates' self-assessments of personal attributes and found that women generally felt more confident than men in possessing nine of the 16 "ideal" attributes listed.[38] These studies indicate that our understanding of students' attitudes, some of which may reflect aspects of "professionalism," can benefit from self-report questionnaires. However, these studies are comparing groups and trends, not assessing the qualities of individuals. The utility of self-reporting for these purposes might be much more severely limited.

Most studies of self-assessment in medicine focus on the assessment of knowledge and skills rather than on professional behavior, but they generally conclude that self-assessment is quite inaccurate.[28,39] If physicians are inaccurate at self-assessment in relatively concrete domains (e.g., knowledge), they are likely to have even greater difficulty in a domain such as professionalism, which is less well defined and more socially value-laden. A recent line of research, for example, introduced a model of self-assessment described as the relative ranking technique, in which each participant ranks a set of skills relative to each other from the skill that needs the most work to the one that needs the least.[40,41] Despite some success as a self-assessment tool in the relatively constrained domain of interviewing skills, the technique was far less useful when applied to residents' self-assessments of the standard components of a ward assessment form. In this context, the authors discovered that although residents were quite willing to say they need "the most work" with their surgical skills, or to improve their knowledge base, all residents responded that they needed "the least work" in colleague and/or team relationships.[41] It appears that when statements are value-laden and abstract (as in issues of professionalism), the bias of social desirability is strong, and self-assessment becomes distorted and potentially misleading.

It is apparent that the use of self-assessment in the evaluation of professionalism is difficult. The methods used do not take context into account, making them somewhat threatening. Perhaps a relative ranking system could be attempted that included only elements of professionalism, such as interpersonal skills, communication skills, respect, and integrity. However, it would still be unlikely for a student to say he or she needs more work with honesty. Again, behaviors rather than abstract definitions would need to be incorporated to overcome this limitation. Until further research is done to better understand the nature of self-assessment, its utility for assessing professional behaviors is likely to be limited to formative evaluations and the setting of personal goals.

Standardized Patients. There is an extensive body of literature on objective structured clinical examinations (OSCEs) and standardized patients (SPs) and their importance in the evaluation of clinical skills. There is no literature specific to the role of either in the evaluation of professionalism or professional behaviors within medicine; however, there are areas in which issues of professionalism and professional behaviors are touched on indirectly.

Using an adaptation of the American Board of Internal Medicine's Physician Satisfaction Questionnaire, Klamen et al. found that SPs could reliably identify some of the professional characteristics of the doctor–patient interaction, including using understandable language and encouraging patients to ask questions.[24,42] By contrast, Schnabel et al. asked SPs to assess empathy, interpersonal skills, and patient satisfaction on a 13-item checklist used in a senior-medical-student OSCE, and found that up to 20 ratings were needed to generate reliable measures.[43] At the extreme, research conducted using OSCE stations to assess students' skills in dealing with ethical issues concluded that 41 stations would be required to achieve good reliability, even if the content domain were narrowed down to one specific ethical dilemma.[44–46]

At least in part, the difficulty with using OSCE scenarios is the ambiguity with which the concepts are defined on the evaluation form. For example, one set of forms used such anchors as "major problems in demeanor or ethical standards resulting in inadequate ability to deal with the patient's problems" and "actions taken may harm the patient."[47,48] In both instances, unacceptable behaviors are not specified, and judgment is left up to the examiner. On a related note, Arnold suggests that the OSCE, as it now exists, does not discriminate between ethical analysis of a problem and communication skills.[49]

Another issue with SPs' assessment is the problem of artificiality. Norman, for example, reported on the experience with a physicians' remediation program that uses standardized patient scenarios.[50] SPs in a simulated office practice, as well as in standard OSCE stations, were asked to rate physicians' interpersonal skills during each encounter. Compared with the office simulations, the OSCE stations had a low reliability and were felt to be "artificial." This may increase the likelihood that students in this setting might act as they *should* rather than as they *would*. On the other hand, one study has reported several professional lapses in the context of a psychiatry OSCE (the most extreme case involving a student's placing a fleeing SP in a headlock for the purpose of restraint).[51] Hodges et al. argue that if stations are more demanding, they may very well discriminate effectively in terms of professional dimensions. Similarly, Vu et al. suggested that SPs' ratings were highly reliable and valid when compared with comments real patients would be expected to make regarding the behaviors they witnessed.[52]

Again, it is apparent that context is important. Methods of assessment that are more true to life may be more useful than those that involve obviously artificial situations. Students may be aware that there is a professionalism station and respond with actions they assume are on the checklists. It would be interesting to include values conflicts in SP scenarios to specifically assess the students' awareness of the professional values that are involved, and to evaluate their responses. In such a case, there may be more than one right answer, so the students' thought-processes about their actions may be more important than the behaviors they actually display. The low reliability of OSCEs, even when limited to specific dimensions of professionalism, is concerning, and many authors have concluded that the greatest utility of this type of assessment may be in the formative evaluation of students.

Longitudinal Observations. More recently, researchers have developed systems for assessing students' professionalism that are triggered by the observation of problematic student behaviors.[1,2,4] The evaluation instrument is a specific form that is completed by a clerkship director or faculty member when a student exhibits unprofessional behavior during a rotation. When more than one form has been completed for a specific student, a meeting between an academic committee and the student occurs and remediation is instituted. These systems are based on the concept that students' professional behaviors must be assessed longitudinally, across numerous clinical rotations. Both studies describing this evaluation tool have been qualitative descriptions of systems that are in place, and further reliability and validity studies are anticipated. Such systems are very promising, despite a lack of rigorous evaluation, and may work well for identifying those students with significant lapses in professional behavior. However, in their present state, they may not prove as useful as a method of evaluating all students. The important advance these authors have made is their acknowledgement that labeling a *student* as "unprofessional" carries a greater negative connotation than simply recording examples of unprofessional *behavior*.

Discussion: Future Directions in the Evaluation of Professional Behavior

It should be apparent from the preceding discussion that evaluating professionalism in medical students and residents has proved to be a difficult task. The definition-driven abstract way of thinking about professionalism creates a dichotomy for faculty: either apply a harsh label, or let the lapse go. We know from previous research that faculty are much more likely to let the lapse go, which effectively suppresses discussion, feedback, and attempts at remediation.[14]

On the other hand, evaluation methods that consider behaviors, rather than individuals, as professional or unprofessional become much less threatening and would be more likely to gain acceptance by faculty and students. The studies reported by Papadakis et al. and Phelan et al. provide two good examples of such systems.[1,2] Perhaps these methods will decrease faculty's reluctance to report behaviors that should lead to remediation; this can only help in promoting students' professional development. As developed, these evaluation forms are intended to identify and document serious lapses in professional behavior, which fortunately occur in only a few students. Future research might focus on ways to make these forms useful in the evaluation of all students. However, it is likely that some barriers to their use would still exist; for example, faculty would still have to decide what constitutes a major or minor infraction. These limitations might be minimized if the behavior is placed in a context (of the person, the situation, the harm caused to others), a fair process of review is used, and reasonable judgment is applied.[53] Then, any decision made would be justifiable and well supported. Arnold and colleagues use a hybrid of the behavioral and abstract in their measurement tool by attaching behavioral descriptors (such as "I have seen residents refer to patients in derogatory terms") to abstract dimensions of professionalism (such as "respectfulness"), which is an interesting potential step in this direction.[54]

We have also argued that professional behavior is much more context-dependent than has usually been acknowledged. All physicians are exposed to situations that challenge their abilities to act professionally, and medical students and residents are no different. In fact, they may be more vulnerable to lapses in professional behavior because of the nature of their training and environment. It is crucial to be aware of the specific context in which a behavior occurs before attempting to evaluate it. For example, Christakis et al. found that the teaching students had received on ethical dilemmas seemed to lack real-life relevance and related more to the context of a practicing physician.[55] Focus groups described different dilemmas, which were unique to a third-year student's experience. They highlighted the conflicts between education, patient care, wanting to be a team player, and fear of a poor evaluation. One overriding feature was the construct of authority: students lack it and are wary of challenging it, which often puts them into conflict.

It may be necessary to study these behaviors in context more closely to determine their frequency and severity. Since we know that faculty, nurses, students, and residents all see different aspects of professionalism in students, it would be important to gain the perspectives of each of these groups in order to be comprehensive. One way could be to involve each of these groups in focus-group discussions, to determine what they consider to be professional and unprofessional behaviors. Their unique perspectives would help in the design of instruments used in all forms of student assessment. Another technique could be to use an anonymous encounter card system to collect information from students, residents, faculty, and nurses, about what behaviors are actually occurring. This may provide us with a more comprehensive set of behaviors on which to base future evaluation methods.

Conflict has also long been identified as a critical component of professional development, and is found as a dominant element in some measures of professional behavior.[9-11] Although such paper-and-pencil instruments are limited by their artificial nature, some researchers have found that professional behavior can best be identified at the time that students are grappling with these conflicts. One potential implication of this finding is that students could be placed in a situation that involves a conflict of values, for example, with a standardized patient. The behaviors the students display, based on the choices they make, could be evaluated. What might be even more informative is an evaluation of the thought process a student goes through to arrive at his or her ultimate choice.

Alternatively, students could be asked to write about professional conflicts they have encountered.[56] The language or text from these experiences could be subjected to linguistic or rhetorical analysis to uncover the underlying values of individual students and explore how these values affect the resolution of professional conflicts. Lin-

20

gard and Haber's studies use a rhetorical framework to explore how the structural patterns of case presentations inform medical students' developing attitudes towards patients and colleagues.[57,58] The authors demonstrate that a rhetorical analysis of discourse patterns can reveal critical relationships between the stories novices learn to tell about patients and the decisions they make about how to act on behalf of and in relation to them. Other studies in a similar vein reinforce the potential usefulness of this method.[59–61] However, the texts that students generate may suffer from the same sense of artificiality that affects OSCE stations, and research in this area would have to be designed to take this issue into account.

It is unrealistic to think that one evaluation instrument could capture all that is important in the complex domain of professionalism. As with all high-stakes evaluations, reliability, which depends in part on sample size, is important. No student should receive a grade on his or her knowledge of cardiology from a single-item test; similarly, no student should receive a grade on professionalism without adequate sampling of the domain. Some of the measures outlined above have large sample sizes and are likely to be more useful (peer evaluation, encounter cards), while others rely on a single report or a few reports (SP scenarios, ward evaluations). While the latter may be useful for outliers, the former are more useful for the larger group of students who experience only occasional lapses in professional behavior. It is certain that more than one measurement technique would need to be used, and the greatest validity may result from triangulating results from different sources.

Future efforts at understanding professionalism, and future methods of evaluating professionalism, must focus on behaviors rather than personality traits or vague concepts of character. Our understanding and evaluation must include context and conflict in order to be relevant and valid. Ideally, methods of evaluation should include elements of peer assessment and self-assessment, which are becoming required elements in the continuing professional development of all practicing physicians. Finally, we should attempt to understand what drives students to demonstrate occasional lapses in professional behavior, in order to develop effective teaching and remediation in this domain.

Correspondence: Shiphra Ginsburg, MD, Mt. Sinai Hospital, 600 University Avenue, Toronto, Ontario, M5G 1X5, Canada.

References

1. Papadakis MA, Osborn MC, Cooke M, Healy K. A strategy for the detection and evaluation of unprofessional behavior in medical students. Acad Med. 1999;74:980–90.
2. Phelan S, Obenshain S, Galey WR. Evaluation of the non-cognitive professional traits of medical students. Acad Med. 1993;68:799–803.
3. Hunt DD, Scott CS, Phillips TJ, Yergan J, Greig LM. Performance of residents who had academic difficulties in medical school. J Med Educ. 1987;62:170–6.
4. Project Professionalism. Philadelphia, PA: American Board of Internal Medicine, 1995.
5. Wolf TM, Balson PM, Faucett JM, Randall HM. A retrospective study of attitude change during medical education. Med Educ. 1989;23:19–23.
6. Feudtner C, Christakis DA, Christakis NA. Do clinical clerks suffer ethical erosion? Students' perceptions of their ethical environment and personal development. Acad Med. 1994;69:670–9.
7. Garfinkel PE, Bagby RM, Waring EM, Dorian B. Boundary violations and personality traits among psychiatrists. Can J Psych. 1997;42:758–63.
8. Graham JR. The MMPI: A Practical Guide. Second edition. New York: Oxford University Press, 1987.
9. Oser FK. Moral education and values education: the discourse perspective. In: Wittrock MC (ed). Handbook of Research on Teaching. New York: Macmillan Publishing Company, 1986:917–41.
10. Stern DT. Hanging out: teaching values in medical education [dissertation]. Stanford, CA: Stanford University, 1996.
11. Rezler AG, Schwartz RL, Obenshain SS, Lambert R, Gibson JM, Bennahum DA. Assessment of ethical decisions and values. Med Educ. 1992;26:7–16.
12. Novack DH, Detering BJ, Arnold R, Forrow L, Ladinsky M, Pezzullo JC. Physicians' attitudes toward using deception to resolve difficult ethical problems. JAMA. 1989;261:2980–5.
13. Stern DT. Practicing what we preach? An analysis of the curriculum of values in medical education. Am J Med. 1998;104:569–75.
14. Burack JH, Irby DM, Carline JD, Root RK, Larson EB. Teaching compassion and respect: attending physicians' responses to problematic behaviors. J Gen Intern Med. 1999;14:49–55.
15. Hunt DD, Carline J, Tonesk X, Yergna J, Siever M, Loebel JP. Types of problem students encountered by clinical teachers on clerkships. Med Educ. 1989;23:14–8.
16. Hunt DD. Functional and dysfunctional characteristics of the prevailing model of clinical evaluation systems in North American medical schools. Acad Med. 1992;67:254–9.
17. Gray JD. Global rating scales in residency education. Acad Med. 1996;71(10 suppl):S55–S63.
18. Woolliscroft JO, Howell JD, Patel BP, Swanson DB. Resident–patient interactions: the humanistic qualities of internal medicine residents assessed by patients, attending physicians, program supervisors, and nurses. Acad Med. 1994;69:216–24.
19. Johnson D, Cujec B. Comparison of self, nurse, and physician assessment of residents rotating through an intensive care unit. Crit Care Med. 1998;26:1811–6.
20. Rhoton MF. A new method to evaluate clinical performance and critical incidents in anesthesia: quantification of daily comments by teachers. Med Educ. 1989;23:280–9.
21. Brennan B, Norman GR. Use of encounter cards for evaluation of residents in obstetrics. Acad Med. 1997;72(10 suppl):S43–S44.
22. Rhoton MF. Professionalism and clinical excellence among anesthesiology residents. Acad Med. 1994;69:313–5.
23. Beckman H, Frankel R, Kihm J, Kulesza G, Geheb M. Measurement and improvement of humanistic skills in first-year trainees. J Gen Intern Med. 1990;5:42–5.
24. Klamen DL, Williams RG. The effect of medical education on students' patient-satisfaction ratings. Acad Med. 1997;72:57–61.
25. Klessig J, Robbins AS, Wieland D, Rubenstein L. Evaluating humanistic attributes of internal medicine residents. J Gen Intern Med. 1989;4:514–22.
26. Matthews DA, Feinstein AR. A new instrument for patients' ratings of physician performance in the hospital setting. J Gen Intern Med. 1989;4:14–22.
27. Butterfield PS, Mazzaferri EL. A new rating form for use by nurses in assessing residents' humanistic behavior. J Gen Intern Med. 1991;6:155–61.
28. Linn BS, Arostegui M, Zeppa R. Performance rating scale for peer and self assessment. Br J Med Educ. 1975;9:98–101.
29. Arnold L, Willoughby L, Calkins V, Gammon L, Eberhart G. Use of peer evaluation in the assessment of medical students. J Med Educ. 1981;56:35–42.
30. Helfer RE. Peer evaluation: its potential usefulness in medical education. Br J Med Educ. 1972;6:224–31.
31. Thomas PA, Gebo KA, Hellmann DB. A pilot study of peer review in residency training. J Gen Intern Med. 1999;14:551–4.
32. Parker PA Jr., Stevens CB, Duerson MC. Issues in medical education: basic problems and potential solutions. Acad Med. 1993;(10 suppl):S89–S93.
33. Ramsey PG, Carline JD, Blank L, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performance of practicing physicians. Acad Med. 1996;71:364–70.
34. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA. 1993;269:1655–60.
35. Van Rosendaal GMA, Jennett PA. Resistance to peer evaluation in an internal medicine residency. Acad Med. 1992;67:63.
36. Flaherty JA. Attitudinal development in medical education. In: Rezler A (ed) The Interpersonal Dimension in Medical Education. New York: Springer, 1985:147–82.
37. Testerman JK, Morton KR, Loo LK, Worthley JS, Lamberton HH. The natural history of cynicism in physicians. Acad Med. 1996;(10 suppl):S43–S5.
38. Clack GB, Head JO. Gender differences in medical graduates' assessment of their personal attributes. Med Educ. 1999;33(2):101–5.
39. Jankowski J, Crombie I, Block R, Mayet J, McLay J, Struthers AD. Self-assessment of medical knowledge: do physicians overestimate or underestimate? J Royal Coll Phys London. 1991;25:306–8.
40. Regehr G, Hodges B, Tiberius R, Lofchy J. Measuring self-assessment skills: an innovative relative ranking model. Acad Med. 71(10 suppl):S52–S4.
41. Harrington JP, Murnaghan JJ, Regehr G. Applying a relative ranking model to the self-assessment of extended performances. Adv Health Sci Educ. 1997;2:17–25.
42. PSQ Project Co-Investigators. Final Report on the Patient Satisfaction Questionnaire Project. Philadelphia, PA: American Board of Internal Medicine, 1989.
43. Schnabel GK, Hassard TH, Kopekow ML. The assessment of interpersonal skills using standardized patients. Acad Med. 1991;66(10 suppl):S34–S36.
44. Singer P, Cohen R, Robb A, Rothman A. The ethics OSCE. J Gen Intern Med.

21

1993;8:23–8.
45. Singer P, Robb A, Horman G, Turnbull J. Performance-based assessment of clinical ethics using an OSCE. Acad Med. 1996;71:495–8.
46. Smith SR, Balint JA, Krause K, Moore-West M, Viles PH. Performance-based assessment of moral reasoning and ethical judgment among medical students. Acad Med. 1994;69:381–6.
47. Reznick RK, Regehr G, Yee G, Rothman A, Blackmore D, Dauphinee D. Process-rating forms versus task-specific checklists in an OSCE for medical licensure. Acad Med. 73;1998(10 suppl):S97–S99.
48. Medical Council of Canada. Information Pamphlet: Qualifying Examination Part II. Ottawa ON: Medical Council of Canada, 1999:11.
49. Arnold RM. Assessing competence in clinical ethics: are we measuring the right behaviors? J Gen Intern Med. 1993;8:52–4.
50. Norman GR, Davis D, Lamb S, Hanna E, Caulfor P, Kaigas T. Competency assessment of primary care physicians as part of a peer review program. JAMA. 1993; 270:1046–51.
51. Hodges B, Turnbull J, Cohen R, Bienenstock A, Norman G. Evaluating communication skills in the OSCE format: reliability and generalizability. Med Educ. 1996;30:38–43.
52. Vu NV, Marcy ML, Verhulst SJ, Barrows HS. Generalizability of standardized patients' ratings of their clinical encounter with fourth-year medical students. Acad Med. 1990;65(10 suppl):S29–S30.
53. Irby DM, Milam S. The legal context for evaluating and dismissing medical students and residents. Acad Med. 64;1989:639–43.
54. Arnold EL, Blank LL, Race KEH, Cipparone N. Can professionalism be measured? The development of a scale for use in the medical environment. Acad Med. 1998; 73:1119–21.
55. Christakis DA, Feudtner C. Ethics in a short white coat: the ethical dilemmas that medical students confront. Acad Med. 1993;68:249–54.
56. Szauter K, Boisaubin E. Professional ethical dilemmas of medical students during the medicine clerkship. Oral abstract presented at the 38th Annual Research in Medical Education Conference, Washington, DC, October 1999.
57. Lingard L, Haber RJ. What do we mean by "relevance"? A clinical and rhetorical definition with implications for teaching and learning the case-presentation format. Acad Med. 1999;74(10 suppl):S124–S127.
58. Lingard LA, Haber RJ. Teaching and learning communication in medicine: a rhetorical approach. Acad Med. 1999;74:507–10.
59. Arluke A. Social control rituals in medicine: the case of death rounds. In: Dingwall R, Heath C, Reid M, Stacey M (eds). Health Care and Health Knowledge. London, U.K.: Croom Helm, 1977:108–25.
60. Caldicott CV. What's wrong with this medical student today? Dysfluency on inpatient rounds. Ann Intern Med. 1998;128:607–10.
61. Stern DT, Caldicott CV. Turfing: patients in the balance. J Gen Intern Med. 1999; 14:243–8.

22

Tracking Knowledge Growth across an Integrated Nutrition Curriculum

CAROL S. HODGSON

Both the academic literature and the popular press continually report the importance of nutrition for health. One example is the increasing prevalence of obesity in the United States[1] and the concomitant popularity of diet books and untested remedies. A 1985 National Academy of Sciences report warned of the lack of nutrition education and the need for a required curriculum for all medical students in U.S. medical schools.[2] Results from annual Association of American Medical Colleges (AAMC) Graduation Questionnaires reinforce this conclusion. In 1995, 63% of students reported that they had received inadequate nutrition in their medical school curricula.[3] In 1998, nothing had changed; 64% of students still reported inadequate nutrition education.[4]

Following the 1985 National Academy of Sciences report, funding from the National Cancer Institute (NCI) stimulated development of nutrition curricula at a number of medical schools.[5-7] In one study, the use of a multimedia program to teach nutritional assessment and counseling was evaluated.[6] The authors found that, following exposure to the multimedia nutrition program, first-year students were more likely to use a food-frequency form while interviewing a standardized patient compared with previous students who had not received the intervention. A majority of students (51%) who completed the curriculum reported that observing a physician model nutritional assessment and counseling in the multimedia program had been helpful. In another study, the evaluation of a two-year integrated nutrition curriculum implemented during the basic sciences indicated increased knowledge for those students who had completed the curriculum.[7] Although limited in scope, these studies are promising. They imply that, even when nutrition is not a major aspect of the medical school curriculum, first- and second-year students' knowledge can increase and they may apply their knowledge to patient care.

The increase of nutrition knowledge following exposure to the clinical curriculum is potentially of greater importance than is the nutritional content of the basic science curriculum, since clinical exposure may be more likely to lead to application in practice. Many studies report physicians' lack of knowledge and confidence in using nutritional concepts in their practices.[8-11] The paucity of physicians who model the use of nutrition concepts in their practices could have a negative effect on students' acquisition of knowledge and their application of that knowledge to patient care.[12]

At our institution, cognitive learning theory (i.e., actively engaging students in learning[13]) guided the development of a new nutrition curriculum. The curriculum's goals were to increase students' (1) learning and retention of nutritional concepts; (2) skills, such as diet-assessment methods; and (3) application of content to patients' care. To accomplish this, we planned to increase opportunities for practice with nutritional concepts throughout the four-year curriculum using active learning methods such as laboratory exercises, a dietary self-assessment, interviews with standardized patients, and discussions in small-group problem-based learning (PBL) sessions.

In 1992, we started the curricular planning process by conducting a nutrition needs assessment. We received funding of an NCI R25 grant (NCI PAR 94-005) in 1994, and a Nutrition Education Committee was formed to develop and implement the new nutrition curriculum. The Committee established goals and objectives (outlined on our Web site (http://apps.medsch.ucla.edu/nutrition/objectives.html)), reviewed existing courses and clerkships, and implemented changes in years one, two and three of the curriculum. New instructional and examination materials were developed to foster accomplishment of nutrition proficiencies outlined on the Web site above. The development of ongoing curricular review and evaluation processes tracked growth of nutritional knowledge in those students exposed to the revised curriculum.

Modification of targeted courses to emphasize proficiency with nutritional concepts was the primary strategy of the curricular change. The nutrition curriculum is concentrated in the first-year course, Human Biochemistry and Nutrition Laboratory. A number of nutrition-related cases are also included in two first-year PBL courses. Nutrition is included in approximately ten lectures of the second-year course, Pathophysiology of Disease. New curricular material was incorporated into the required third-year family medicine clerkship and the Doctoring 3 curriculum, where students interview standardized patients. Numerous fourth-year nutrition electives are offered, but their impact is limited because very few students take these electives.

In this study, we examined the effect of changes in the nutrition curriculum on students' knowledge over four years of medical school. Based on earlier findings, and further development and implementation of nutritional content in the clinical curriculum, we hypothesized that students completing an integrated four-year nutrition curriculum would demonstrate, on a Nutrition Progress Survey, a continual increase in their nutrition knowledge over time. We also hypothesized that they would demonstrate more confidence in their responses through a decrease in their use of "don't know" as a response to survey questions.

Method

We used a pre-/post-test intact-group design to evaluate changes in the nutrition knowledge of a cohort of medical students as they progressed from their first to fourth years (class of 1998). Test items that originally had been developed at the University of Alabama and had been demonstrated to be valid and reliable measures for assessing the nutritional knowledge of medical students formed the basis of our 90-item Nutrition Knowledge Progress Survey. In order to decrease the use of guessing, students were given an additional response option, "don't know," for all questions. The students were informed that the test items would be scored (correct = +1, incorrect = −1, and don't know = 0). All students completed an informed consent form prior to entering the study.

The nutrition survey was administered as a pre-test to the first-year class in January 1995. A 45-item subtest of the survey (30 items expected by first-year course chairs to be initially covered in the first-year curriculum and 15 randomly selected items) was administered in May 1995 (post-test 1) to the same cohort of students. Delayed post-test exams were given to third-year students in August 1996 (post-test 2) and to fourth-year students in August 1997 (post-test 3). Two forms of post-test 2 were administered to third-year students: the full 90-item and the 45-item subtests of the survey. Earlier we reported no significant difference between the scores on the 45 items in common on the two forms of the test.[7] The 90-item exam was given at the start of the fourth year to a randomly selected half of the cohort (n = 76). Those students who completed all four previous surveys were asked to complete one more at the end of the fourth year (post-test 4). Each fourth-year

323

TABLE 1. Students' Responses to the Nutrition Knowledge Progress Survey*

	Pre-test Mean (SD)	Post-test 1 Mean (SD)	Post-test 2 Mean (SD)	Post-test 3 Mean (SD)	Repeated-measures ANOVA† F (df)	Repeated-measures ANOVA† $p <$
Total score	9.50 (5.40)	15.25 (7.88)‡	18.21 (5.80)§	22.96 (5.38)¶	37.03 (3, 21)	.001
Number correct	18.04 (4.67)	26.50 (4.29)‡	28.79 (4.41)§	32.04 (3.57)¶	61.89 (3, 21)	.001
Number "don't know"	18.42 (5.36)	7.25 (4.24)‡	5.62 (4.82)§	3.50 (3.04)¶	64.69 (3, 21)	.001
Number incorrect	8.54 (2.67)	11.25 (4.65)‡	10.58 (2.99)	9.08 (2.65)¶	3.70 (3, 21)	.05

*Total score (scored +1 for a correct answer, 0 for "don't know" and −1 for an incorrect answer), number correct, incorrect, and "don't know" for 45 items in common administration for those students who completed the first four test administrations (n = 24) compared across four test administrations using a repeated-measures ANOVA.
†Significant contrasts using difference method are: ‡comparing post-test 1 with pre-test; §comparing post-test 2 with post-test 1; ¶comparing post-test 3 with post-test 2.

student who participated received a $100 gift certificate as an incentive.

Total scores were calculated by summing the scores for the items in each exam (correct = +1, incorrect = −1, and don't know = 0). The 45 items in common for each test were summed to form a total score for each administration of the survey. Additionally, the 30 items in the survey covered in the first-year curriculum and used in subsequent years were summed to create a total score for the first-year curriculum in order to test for learning and retention of material. In order to test whether the total numbers of correct, incorrect, and "don't know" answers changed over time, total scores for these answers were calculated by summing the number of responses for each category. A repeated-measures analysis of variance (ANOVA) was used to examine changes between the time points. The difference method was used to compare each time point with the previous one.

It was possible that those students who completed the survey every time it was given differed from those students who did not (i.e., were more knowledgeable about nutrition). In order to test this, a sample was randomly selected (equaling the sample size of those who completed all four surveys) from those students who had not completed all four exams. Pre-test, post-test 1, and post-test 2 total scores were compared for these two groups.

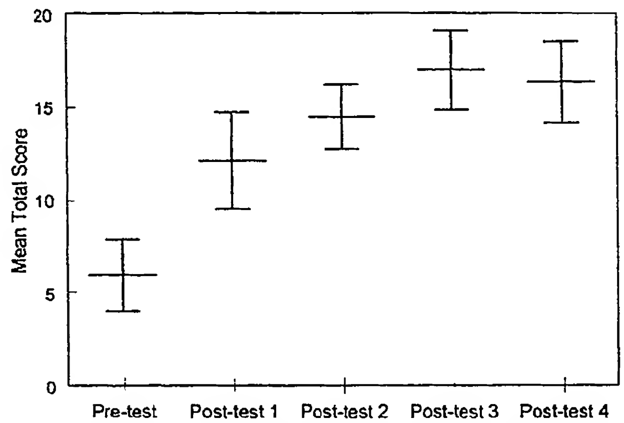

Figure 1. Mean scores ± 2 standard errors on the Nutrition Knowledge Progress Survey (30 items from the first-year curriculum) for those students who completed all five test administrations (n = 20). Repeated-measures ANOVA: f = 23.3 (4, 16), $p < .001$. The test scored +1 for correct response, 0 for "don't know," and −1 for an incorrect answer.

Results

Approximately 90% of the cohort completed at least one of the four exams: 88% at pre-test (n = 130), 93% at post-test 1 (n = 136), 72% at post-test 2 (n = 89), and 70% at post-test 3 (53 of 76 students recruited to complete the nutrition survey). Fifty-three percent of the students (n = 78) completed the first three exams. Twenty of the 24 students (83%) who filled out the first four surveys completed post-test 4.

The first set of data reported includes all students who completed the nutrition survey at the first four test administrations except for the comparison group. The second set of data reported includes only those students who completed the survey at all five administrations.

There was a significant increase in knowledge over the four test administrations (see Table 1). Results of the repeated-measures ANOVA showed a significant increase in knowledge over the three-year time period using the 45-item subtest. The number of correct answers increased; the numbers of incorrect and "don't know" responses decreased. Within-subject comparisons between each time period and the previous time period were also significant (see Table 1), indicating a significant increase in knowledge from one time point to the next. In addition, knowledge relative to the content covered in the first-year curriculum (30-item subtest) increased over time (see Figure 1).

Students who completed the first four nutrition surveys (n = 24) were compared with randomly selected groups of students who did not complete all four surveys on pre-test, post-test 1, and post-test 2 mean scores. There was no significant difference between the two groups on any of these measures, indicating that there was no bias in terms of nutrition knowledge as to who completed all of the four nutrition surveys.

Discussion

Results from this study indicate that the goals of the curriculum were met; medical students who received the longitudinal integrated nutrition curriculum did increase their knowledge over time and retained the knowledge learned in the first year through the third year (see Figure 1). In addition, students appeared to be more confident in their responses, since they decreased their use of the "don't know" response, even though they risked losing points for an incorrect answer. These results, however, do not mean that students are more able to apply their knowledge in the clinical setting. In contrast, anecdotally, we know from speaking informally with fourth-year medical students that they felt very uncomfortable being alone in an exam room with a patient who asked about diet or supplements. Results from the AAMC Graduation Questionnaire confirmed this. Even though our students clearly increased their knowledge over time, 68% of this cohort still reported inadequate education in nutrition in their curriculum, compared with 64% nationally.[4] This finding might reflect the students' greater understanding of the importance of nutrition in clinical practice based on the curriculum. On the other hand, it may be that their own clinical experience, although limited, had informed them of their need to know.

24

There are a number of limitations to this study. First, only one school was studied, so results might not be comparable in another school. There may have been a test effect from using the same survey over time, although given the time lag between administrations and the lack of grading associated with it, this seems unlikely. There may have been sample bias if those students who completed the survey all five times were more interested in nutrition. Again, this is unlikely given the comparison of those students who took all five tests with those who took only the pre-test, post-test 1, or post-test 2. Finally, it is possible that the results are purely from a maturation effect. This is not likely, however, given our earlier study results indicating no significant difference in a comparison of scores on post-test 2 of a control group (those not completing the nutrition curriculum) with scores of students who had completed the nutrition curriculum.[7]

Results from this study are promising, but there is still a way to go—one of the biggest hurdles remaining is incorporating nutrition into the clinical curriculum. The average number of items answered correctly by those graduating students who completed the survey was 32 of 45, indicating a marginally passing score of 71%. This denotes an increase of only 13% from their scores at the end of the first year. However, these results are similar to those of a multi-school study conducted in the late 1980s, in which fourth-year students at 11 southeastern U.S. medical schools scored an average of 69% on a similar survey. Scores were related to the amount of required nutrition curriculum the students had experienced. Although knowledge scores increased, students' attitudes with respect to the importance of nutrition for their careers deteriorated from year one to the end of the clinical curriculum.[14] At our institution, nutritional content increased in the third-year curriculum, but little advancement was made into any clerkship except family medicine. Given the general lack of nutrition knowledge of clinicians,[8-11] it is likely that there were few preceptor role models who demonstrated or reinforced nutritional assessment or dietary counseling of patients. Consistent with this are the results of a study comparing the nutrition knowledge of our fourth-year students with that of physicians attending a local nutrition continuing medical education course. The students significantly outscored the physicians in nutrition knowledge (68% versus 52%).[15]

Last, although a case with nutritional content was inserted into our senior clinical performance examination, this occurred after this cohort of students had graduated. Therefore, the only change observed, (an increase of students' knowledge of nutrition concepts) provides no evidence that students will apply this information in clinical practice—our ultimate goal. Further studies are needed to examine this potential effect of the curriculum.

This work was supported by a National Cancer Institute R25 grant (NCI PAR 94-005).

Correspondence: Carol S. Hodgson, PhD, UCLA School of Medicine, 10833 Le Conte, 60-051 CHS, Los Angeles, CA 90095-1722.

References

1. Must A, Spadano J, Coakley EH, et al. The disease burden associated with overweight and obesity. JAMA. 1999;282:1523–9.
2. National Research Council. Nutrition education in US medical schools. Washington, DC: National Academy Press, 1985.
3. Association of American Medical Colleges. 1995 Medical School Graduation Questionnaire Survey Results: All Schools Summary. Washington, DC: AAMC, 1995.
4. Association of American Medical Colleges. 1998 Medical School Graduation Questionnaire Survey Results: All Schools Summary. Washington, DC: AAMC, 1998.
5. Johnson VK, Murphy G, Michener JL. An integrated nutrition curriculum for medical students. Acad Med. 1995;70:433–4.
6. Kolasa KM, Jobe AC, Clay M, Daugherty J. Evaluating the use of a multimedia approach to teaching nutrition in medical school. J Nutr Educ. 1997;29:351–5.
7. Hodgson CS, Wilkerson L, Go VWL. Changes in nutrition knowledge among first- and second-year medical students following implementation of an integrated nutrition curriculum. J Cancer Educ. 2000;15:144–7.
8. Francis J, Roche M, Mant D, et al. Would primary health care workers give appropriate dietary advice after cholesterol screening? BMJ. 1989;298:1620–2.
9. Hiddink GJ, Hautvast JG, Van Woerkum CM, et al. Nutrition guidance by primary-care physicians: perceived barriers and low involvement. Eur J Clin Nutr. 1995;49:842–51.
10. Kolasa KM. Developments and challenges in family practice nutrition education for residents and practicing physicians: an overview of the North American experience. Eur J Clin Nutr. 1999;53, suppl 2, S89–S96.
11. Tziraki C, Graubard BI, Manley M, et al. Effect of training on adoption of cancer prevention nutrition-related activities by primary care practices: results of a randomized, controlled study. J Gen Intern Med. 2000;15:155–62.
12. Halstead CH. The relevance of clinical nutrition education and role models to the practice of medicine. Eur J Clin Nutr. 1999;53(S2):S29–S34.
13. Gagne ED. The Cognitive Psychology of School Learning. New York: HarperCollins College Publishers, 1993.
14. Weinsier RL, Boker JR, Morgan SL, et al. Cross-sectional study of nutrition knowledge and attitudes of medical students at three points in their medical training at 11 southeastern medical schools. Am J Clin Nutr. 1988;48(1):1–6.
15. Hodgson CS, Wilkerson L. Comparison of nutrition knowledge of physicians attending nutrition CME with medical students. Presented at the annual meting of the American Association for Cancer Education, Portland, OR, October 1998.

25

Following Medical School Graduates into Practice: Residency Directors' Assessments after the First Year of Residency

GWEN L. ALEXANDER, WAYNE K. DAVIS, ALICE C. YAN, and JOSEPH C. FANTONE III

Extensive resources are devoted to preparing medical students to practice in the demanding world of medicine. While students' progress is extensively monitored during medical school, very few medical schools have reported research showing the relationship of medical school preparation to performance during residency education.[1-4] There is growing recognition of the need for measurable outcomes of medical education. Performances of graduates in their residency programs provide one outcome that could be used to assess the quality of medical school educational programs. The purpose of this study was to consider information about the performances of our graduates, assessed early in their residency education by residency program directors, and to explore the relationship between those ratings and our graduates' performance evaluations during medical school.

In the spring of 1997, the University of Michigan Medical School (UMMS), in Ann Arbor, Michigan, began a longitudinal follow-up program designed to collect residency directors' assessments of the performances of our graduates at the end of their first year of residency. This investigation was, in part, inspired by the Liaison Committee on Medical Education (LCME) statement that medical schools must evaluate the effectiveness of educational programs and document graduates' achievement, showing the extent to which institutional and program purposes are met.[5] Initiation of this study coincided with the completion of an extensive and incremental curricular change. The goals of the curricular change, reflecting changes in educational goals, included more opportunities for clinical applications of medical science and hands-on, active learning in the first two years. Extensive efforts were made to encourage collegiality and professionalism among students, and more frequent and earlier patient encounters to promote a more humanistic, patient-centered approach to medical decision making. The evaluation system was also revised to pass/fail grading in the first year, with additional mechanisms implemented to ensure earlier and increased feedback to students from objective measures throughout the first two years of medical school.

An important goal of this research project was to validate the system used to assess students' performances in medical school by comparing the medical school's assessments with performance assessments of UMMS graduates early in their residency education. In particular, we wanted to assess the contributions of academic assessments at various intervals during medical school to ratings of residency performance across all students and by subgroups based on academic achievement, gender, and ethnicity.

Method

To collect residency directors' ratings of our graduates' skills and abilities, we developed an instrument representing various domains of medical practice and aligned with the key goals of our revised curriculum. The seven domains included in the instrument were clinical judgment, patient management, clinical skills, professional qualities, humanistic qualities, oral and written presentation skills, and a final overall performance assessment question. The survey instrument used a five-point Likert-type response format (1 = poor, 2 = fair, 3 = good, 4 = very good, and 5 = excellent). Residency directors were also asked to make written narrative comments on the instrument. Our intention was to construct an instrument that would be self-explanatory and that could be completed in five minutes or less. Curriculum committee members approved the finalized survey.

The survey was mailed to the residency directors of the UMMS graduating classes of 1996, 1997, and 1998 in May of the graduates' first year of residency. Responses were categorized by each graduate's residency specialty type and by his or her program's affiliation with either a community-based or a university-based hospital.

Medical school assessments considered in the analyses included overall grade-point average (GPA) of the second medical science year (M2), U.S. Medical Licensing Examination (USMLE) Step 1 scores, overall grade-point average of the seven required clerkships in the third (clinical) year (M3), USMLE Step 2 scores, and a cumulative composite score at graduation. This composite score at graduation was composed of a grade-point average computed over all second-, third-, and fourth-year courses, with a small fraction representing USMLE Step 1 and Step 2 scores, using the formula of medical school cumulative grade-point average (GPA) + [(USMLE 1 + USMLE 2)/4,000].

The structure of the instrument was assessed using principal-components factor analysis. Cronbach's alpha was used to determine the instrument's internal consistency. Responses were initially analyzed by graduating class. Demographic, program, and academic achievement variables were compared to determine representativeness of responses. Descriptive statistics for the individual items on the survey were compared based on the residency program's affiliation, specialty subgroup, and the gender of the graduate. A lack of differences among individual graduation years allowed the combination of data from all three years. Correlations were computed between measures of medical school performance and directors' ratings. A one-way analysis of variance (ANOVA) was used to compare subgroup means, utilizing post-hoc tests for mean differences.

Results

A single mailing of the survey instrument was sent to residency directors of 498 graduates of three consecutive graduating classes, and 338 (68%) were returned. The graduates represented by directors' responses were 61% men and 39% women. The residents' racial–ethnic subgroups were Asian (16%), underrepresented minority (15%), and white and all others (69%). The residents' specialty subgroups were primary care (50%), surgery and surgery subspecialties (27%), and all other specialties (23%). The 136 graduates not represented by directors' responses were statistically similar in distribution by gender, ethnicity group, average overall M2 GPA, average overall M3 clerkship performance GPA, and average USMLE Step 1 scores. The return rate from directors of surgery subspecialties was lower than those of other residency specialty groups (chi-square = 10.2, $p < .002$).

Across all responses, the average ratings for individual survey items were above 4.0 (very good), with the highest average ratings given for the items assessing humanistic and professional qualities. Although several content areas were included in the instrument, factor analysis of the domains represented by the instrument's seven items demonstrated a single factor, explaining 74% of the variance in scores. Internal consistency of the items in the single factor was high (Cronbach alpha of .94). These findings suggested that the

26

Table 1. Pearson Correlations* between University of Michigan Medical School Performance Evaluations and Overall Performances Assessed by Residency Directors for the Graduating Classes of 1996, 1997, and 1998

	M2 GPA	USMLE Step 1	M3 GPA	USMLE Step 2	Overall Graduation Composite Score
USMLE Step 1	.82				
M3 GPA	.63	.58			
USMLE Step 2	.69	.81	.64		
Overall graduation composite score†	.84	.72	.85	.70	
Overall performance‡ (residency)	.20	.20	.41	.24	.32

*All Pearson correlations significant ($p < .000$), $n = 338$.

†Overall graduation composite score is composed of a grade-point average computed over all second-, third-, and fourth-year courses with a small fraction representing USMLE Step 1 and Step 2 scores, using the formula medical school cumulative GPA + [(USMLE Step 1 + USMLE Step 2)/4,000].

‡Overall performance is a single item representing the seven domains included in the survey completed by residency program directors.

Table 2. Comparisons of Residency Directors' Mean Assessments of Overall Performance by University of Michigan Medical School Grading Components, Classes of 1996, 1997, 1998*

Undergraduate Performance Measures	Performance Level and Directors' Ratings		
	Lowest Third Mean (SD)	Middle Third Mean (SD)	Top Third Mean (SD)
M2 grade-point average	4.08 (0.81) $n = 107$	4.14 (0.79) $n = 97$	4.42† (0.76) $n = 101$
USMLE Step 1 score	3.97 (0.79) $n = 107$	4.34 (0.68) $n = 104$	4.32‡ (0.84) $n = 117$
M3 grade-point average	3.80 (0.92) $n = 98$	4.12 (0.73) $n = 125$	4.61§ (0.58) $n = 111$
USMLE Step 2 score	4.04 (0.81) $n = 114$	4.17 (0.75) $n = 110$	4.39† (0.75) $n = 113$
Overall graduation composite score	3.94 (0.88) $n = 107$	4.15 (0.74) $n = 118$	4.50† (0.73) $n = 112$

*Using this table, for example, students whose medical school GPAs were in the lowest third received mean ratings of 4.08 from their residency directors, those receiving GPAs in the middle third received mean ratings of 4.14, and those with the highest GPA received mean ratings of 4.42. Mean ratings from residency directors were based their responses to a summary item "overall performance" on an 8-item questionnaire using a 5-point Likert-type rating scale (1 = poor, 5 = excellent).

†Top third differs from lower third and middle third, $p < .05$.

‡Lower third differs from middle and top third, $p < .05$.

§All groups differ, $p < .05$.

survey was measuring the directors' singular perceptions of the residents' performances. A decision was made to use the instrument's final item, "overall performance," to represent directors' assessments in all further analyses.

Inter-item correlations were high ($p < .000$) between the individual grading indices during medical school (M2 GPA, M3 GPA, USMLE Step 1 scores, USMLE Step 2 scores, and overall cumulative composite score). (See Table 1.) The correlation between M3 clinical grades and the overall performance item assessed by program directors was stronger ($r = .41$) than that between the composite cumulative grade at graduation ($r = .32$). The correlation between M3 grade average and overall residency performance rating was nearly twice the magnitude of the correlations between M2 overall GPA, USMLE Step 1, or Step 2 scores and the overall residency performance. (See Table 1.) When we looked at the inter-item correlation between the various medical school assessment components and the seven individual domains of our instrument, we found the relationships to be positive and statistically significant for all individual domains except one; humanistic qualities, assessed by residency directors, was not related to overall M2 grades ($r = .07$, $p = .12$).

Another analysis examining the relationship of undergraduate medical school grades to assessments of residency performance compared subgroups composed of thirds of the class, based on an overall composite score at graduation. Performance of graduates who had been in the top third of their class, on average, was rated higher than was performance of those who were in the lowest third of the graduating class (see Table 2). This relationship held when comparing top and lower thirds based on all medical school assessment components considered in our study (M2 overall GPA, USMLE Step 1, M3 overall GPA, and USMLE Step 2 scores.) The greatest differences between groups were found when comparing thirds of the class based on M3 overall GPA. Statistically significant differences were found when comparing directors' mean ratings of overall performances between those in the lowest and middle thirds, and again when comparing the middle third's with the top third's average ratings ($p < .05$).

Comparisons of our graduates' ratings by gender, by residency specialty (grouped by primary care, surgery and surgery subspecialties, and all other subspecialties), and by residency program affiliation (either community-based or university-based residency programs) showed no difference, on average, for overall residency performance. When the race–ethnicity of graduates was considered, using the three subgroups of underrepresented minority students, Asian students, and all other students, no difference was found in the comparison of program directors' ratings of overall residency performance with cumulative composite scores at graduation. Regardless of their racial–ethnic subgroups, the students in the top third of the class, based on the cumulative composite score, were rated higher by program directors than were the students in the lowest third.

Discussion

Concerns expressed about the participation rates of residency directors at the onset of this project were dispelled. Our relatively high response rate without follow up is consistent with other researchers' efforts,[1,4] and it provides evidence that residency directors are willing to provide assessments of graduates' performances and feedback to medical schools regarding graduates.

Finding that the survey measured essentially one dimension of our graduates' early residency performance was consistent with findings of other studies.[4] Unlike our medical school's composite index, which was computed from many individual measures, the program directors were providing ratings on single items. It is possible that the residency directors based their ratings on a single overarching impression of our graduates that spilled over into ratings of performance in all domains, rather than making distinctions of the strengths and weaknesses of individuals.[6] Just as our medical school combined performance measures across multiple courses and learning experiences in an "overall" percentage of GPA index for our students, the residency directors in our study tended to make global assessments of the residents' performances rather than distinctions among the items in the survey.

We were encouraged that our graduates were rated, on average, as "very good" or higher by residency directors. The consistency of our graduates' ratings, across specialty areas and regardless of university- or community-based program affiliation, provided confirmation that our graduates are prepared and adaptable to medical practice in a variety of settings.

As expected, positive and relatively high correlations were found among the grading components during medical school (M2 GPA, M3 composite scores, cumulative composite score, and USMLE

27

Step 1 and Step 2 scores). While low in magnitude, the correlations between residents' medical school grades and their residency directors' assessments were statistically significant. Although these findings support the relationship between medical school achievement and later performance, academic performance in this study explained less than 20% of the variance in overall residency performance. Academic assessments of this type in medical school do not appear to be capturing other important factors contributing to directors' assessments after graduation.[3,4] The strength of the correlation between the M3 GPA and the residency directors' assessments may have been due to a "method effect" of the ratings provided.[6] Just as the residency directors made largely subjective assessments of our graduates, the majority of the overall clerkship grades for the required clerkships are provided by attendings' ratings of students' clinical performances. It is possible that the number of students in a residency program and the degree of familiarity between the residency director and the graduate may have contributed to the rating patterns.

Combining data across three years achieved an "n" large enough to compare a variety of subgroups. We found that the students represented in the top thirds of their classes, for all academic measures in this study except the USMLE Step 1 scores, were rated higher by residency directors, on average, when compared with the students in the middle and lowest thirds of their classes. Average ratings based on thirds of the class by clerkship performance in the M3 year proved to be the most consistent with the residency directors' average ratings of our graduates' performances. Further, subgroup analyses showed that medical school performance accounted for the difference in program directors' ratings, regardless of a graduate's gender or race–ethnicity.

While it may be intuitive that quality of performance after medical school relies on quality of achievement before graduation, our findings provide evidence to support this. We were able to demonstrate a correspondence between students' performances in components of our school's evaluation system and residency directors' ratings of their subsequent performances. The relationship we found between academic achievement during medical school and performance in residency lends validity to the evaluation system utilized by our medical school, and supports the use of these postgraduate outcomes as measurements of educational programs. Identifying standardized, objective measures that could be utilized as an index of residency performance, similar to those used by our medical school evaluation system, might enhance the value of residency performance ratings as an educational outcome.

The findings of this study are additionally important in increasing our understanding of factors that do not appear to contribute to performance ratings in graduate education. Based on our data, specialty type, gender, and race–ethnicity of graduates, when academic achievement was taken into account, were not contributing factors in residency performance ratings. Discovering and measuring contributing factors other than those included in our evaluation system is our challenge in medical education.

Correspondence: Gwen L. Alexander, PhD, University of Michigan Medical School, Department of Medical Education, G1211 Towsley Center, Ann Arbor, MI 49109-0201.

References

1. Hojat M, Borenstein BD, Veloski JJ. Cognitive and noncognitive factors in predicting the clinical performance of medical school graduates. J Med Educ. 1988;63:323–5.
2. Blacklow RS, Goepp CE, Hojat M. Further psychometric evaluations of a class-ranking model as a predictor of graduates' clinical competence in the first year of residency. Acad Med. 1993;68:295–7.
3. Dawson-Saunders B, Paiva REA. The validity of clerkship performance evaluations. Med Educ. 1986;20:240–5.
4. Yindral KJ, Rosenfeld PS, Donnelly MB. Medical school achievements as predictors of residency performance. J Med Educ. 1988;63:356–63.
5. Liaison Committee on Medical Education. Functions and structure of a medical school: Accreditation and the Liaison Committee on Medical Education—Standards for accreditation of medical education programs leading to the M.D. Degree. Washington, DC: Association of American Medical Colleges, 1998, 13.
6. Hull AL, Hodder S, Berger B, et al. Validity of three clinical performance assessments of internal medicine clerks. Acad Med. 1995;70:517–22.

28

The Impact of an Alternative Approach to Computing Station Cut Scores in an OSCE

JODI HEROLD McILROY

The OSCE is gaining widespread recognition as a valid means of assessing entry-to-practice competence, or eligibility for licensure, of physicians, physiotherapists, and other health professionals. Given the high-stakes nature of these licensure OSCEs, robust psychometric properties of the exams are essential. One of these properties is the resistance of cut scores used in determining pass–fail decisions to such sources of error as differences in examiner perceptions of competence and examiner stringency in judging competence.

A number of standard-setting methods have been described in the literature on performance-based assessment. Methods are typically categorized as relative or absolute,[1,2] with most administrators responsible for high-stakes examinations preferring absolute or criterion-referenced methods. Absolute standard-setting methods compare candidates' performances with an externally determined or defined measure (criterion) and are typically categorized as test-centered or examinee-centered.[1,2] These categories distinguish methods according to whether judgments about competence are based primarily on inspection of test items (e.g., Angoff, Ebel, and similar methods) or on judgments about examinees (e.g., contrasting groups, borderline group). The common elements for all methods include (1) use of expert judges and (2) reference to a hypothetical "minimally competent" person or a hypothetical "borderline competent" performance.[1] Descriptions and classifications of standard-setting methods can be found in review articles by Cizek,[1] Berk,[3] and Cusimano.[4]

A modification of the mean-borderline-group method that is now being employed by a number of credentialling agencies entails identification of a subgroup of candidates actually performing the exam who are identified by the examiners as having a level of clinical competence that is just on the borderline between being competent and not being competent. The station scores for this borderline group are averaged to generate the station cut score. In this approach, the rating of candidates' performances as competent, borderline, or not competent is concurrent with completion of checklists and/or other scoring rubrics by these same examiners.

This modified mean-borderline-group method has potentially interesting implications for the determination of cut scores when large-scale, multi-site examinations are employed. The cut score is calculated as the mean of the scores of all candidates who receive borderline ratings, regardless of site of administration (or examiner). Thus, in a multi-site examination where examiners are nested within sites, an examiner who identifies a greater number of "borderline competent" candidates during the exam has a greater influence than other examiners on the resultant cut score for that station. The inequality of examiners' influence over station cut scores is inconsistent with other standard-setting methods described, such as the Angoff and Ebel methods, and could be problematic when combined with the potential for examiners' differing perceptions of what constitutes a borderline performance.

The proposed alternative to this current approach is one in which every examiner's opinion or concept of borderline competence is weighted the same. In other words, the mean of each examiner's borderline group is calculated first, then the mean across examiners evaluating the same station is calculated to determine the cut score. The impacts of individual examiners on the resultant cut score are thus equalized. The effect of the alternative method on cut scores and the practical impact on pass-fail decisions was explored in order to determine whether further investigation of cut score validity is required.

Method

Data for 1,373 candidates who participated in four administrations (years) of an OSCE used in a national physiotherapy examination were used in the study. Each administration of the OSCE consisted of 20 stations in which candidates were required to perform a clinical skill in the context of a clinical scenario. Results from two stations had been removed for administrative reasons, so these results were not included in the data set provided. Therefore, the scores for a total of 78 stations were used in the study.

Candidates rotated through circuits of ten ten-minute stations and ten five-minute stations. Each site consisted of two ten-minute circuits per five-minute circuit, or two examiners of each ten-minute station, and one examiner of each five-minute station. Each year had a median of 40 candidates per site. (Therefore, as a rule, ten-minute station examiners evaluated 20 candidates and five-minute station examiners evaluated 40 candidates.)

There were seven, five, 12, and 14 sites of administration in the four respective years of the exam. Each candidate was allowed to choose the sites at which he or she participated, so assignment of candidates to sites was not a random process and will likely have been influenced by location of training. The examiners were clinicians from the local community. They were assigned to stations according to their self-identified areas of clinical expertise. They attended a training session where they oriented to the examination procedures and scoring processes before the exam.

Scoring of the OSCE. The candidates performances were rated using a task-specific dichotomous checklist where clinician examiners record whether key behaviors are demonstrated correctly.* In addition, overall performance was rated on a six-point rating scale. The two middle anchors (3 of 6 and 4 of 6) on this scale were "borderline unsatisfactory" and "borderline satisfactory." The borderline group used in computation of cut scores is considered to include all candidates assigned either of these two scores. The overall rating of performance was considered for the sole purpose of identifying borderline candidates to calculate cut scores.

Computation of Cut Scores. The traditional approach entailed finding the mean checklist score for all candidates identified as borderline, regardless of site or examiner. The alternative approach entailed finding the examiner-specific mean checklist score for the borderline group, then computing the average of these means across examiners. This second method, in effect, weights all examiners' opinions equally. These two checklist-based cut scores were computed for the 78 stations used over the four examinations.

In addition to the overall examination score, candidates are required to perform satisfactorily in a criterion number of stations to pass. The number of stations required to pass an examination fluctuates from year to year, depending on the level of difficulty of the examination. In the hypothetical situation constructed for the purpose of these analyses, I used 12- and 13-station cri-

*The cut score is, in practice, applied to a station score that is a composite of a number of scores assigned to different aspects of candidate performance. For the sake of simplicity, the study considered only checklist scores in the analyses.

teria to examine two different scenarios for the impact of using the alternative method of computation on exam-level pass–fail decisions.

Results

When I examined descriptive data for examiner patterns with regard to use of the "borderline competent" rating, the findings were very consistent for the five-minute and ten-minute stations. For all analyses presented, the results are pooled across the 78 stations to maximize power.

There were observed differences in the proportions of candidates deemed borderline by different examiners examining the same station. The mean discrepancy (the range in the proportions of candidates identified as borderline by different examiners of a given station) across the 78 stations was 48%, with the lowest discrepancy being 11% (where an examiner at one site identified no candidate as borderline, and an examiner at another site identified 11% as borderline) and the highest discrepancy being 90% (with one examiner identifying no borderline candidate and another identifying 90% of candidates as borderline). Clearly, in the computation of cut scores, some examiners are contributing substantially more borderline candidates than others.

The examiner-specific cut scores (the mean checklist score for borderline candidates at a site) also displayed within-station ranges. For example, there was a cut-score discrepancy of 56% (of total possible checklist points) between two examiners of a given station on two of the 78 stations examined. Thirty-five of the 78 stations (45%) had cut score ranges of 30% or more across examiners.

Thus, when discrepancies in examiner-specific cut scores were considered in conjunction with discrepancies in the proportions of candidates rated by individual examiners as borderline there was high potential that, in this hypothetical case based on checklist scores only, use of the proposed alternative method of computation could lead to very different results, at least at the level of station cut scores and pass–fail decisions. The remaining analyses assessed the impact of the observed variability among examiners in their applications of the borderline rating on exam results.

Impact on Raw Cut Scores. The two computation methods generated very similar ranges of station cut scores across the 78 stations. The traditional method resulted in cut scores ranging from 37.03% for the most difficult station to 87.13% for the easiest, while the alternative method resulted in cut scores ranging from 35.37% to 86.93%. At the level of station, differences between the two cut scores were strikingly small, with the maximum difference in cut scores between the two methods being 4.73%. Differences between cut scores were relatively normally distributed around a mean difference of 0.31. There was no significant difference between the cut-scores using the two different approaches (paired t_{77} = 1.66, p = .10).

Also worth noting is the fact that of the 78 stations, 22 were used on more than one occasion. From these stations, it was possible to assess the stability of cut scores over time. The absolute difference in cut scores for the two iterations of each station was calculated for each method. The two methods showed equal levels of cut-score stability of the 22 stations over multiple occasions, in that there was no significant difference in the absolute difference score (reflective of cut score change over time) between methods (paired t_{21} = 1.01, p = .33).

Impact on Station-level Pass–Fail Decisions. There was an equally small effect on pass–fail decisions made at the level of station. For 58 of the 78 stations (74%) there was perfect concordance of pass–fail decisions made using the two methods (i.e., failure rates were unaffected by use of the alternative method). For the 20 stations that were affected, the alternative method increased failure rates at 13 stations while decreasing rates at seven stations. Changes in failure rates for the 20 affected stations ranged from 2% to 18% of candidates examined. There was, however, no significant difference in station failure rates between methods (paired t_{77} = 0.78, p = .44).

When the 22 repeat-use stations were examined for stability of station-level pass–fail decisions, the alternative computation method did not result in a substantial practical effect. The absolute differences between failure rates at two occasions of station use were no different when compared for the two methods (paired t_{21} = 0.24, p = .82).

Impact on Exam-level Pass–Fail Decisions. The examination-level pass–fail decisions are made on the basis of the number of stations passed. In other words, their station scores must be above the station cut score on, say, 12 of 20 stations. In the hypothetical situation created for the study, candidates were required to meet a criterion of passing 12 or 13 stations (both scenarios were examined), based on historical precedent at this and other similar testing organizations. When candidate performance data were used to examine whether the effect of using the alternative cut-score computation method on station-level decisions would translate into an effect at the level of the entire examination, very little impact was seen. The exam-level agreement rates for the two methods (i.e., the proportion of candidates where the exam-level pass–fail decision was unaffected) ranged from 95% to 98% for the four exams, with kappa coefficients ranging from 0.88 to 0.96. Pooled across all four years of administration, the agreement rates were 96% for a 13-station criterion and 97% for a 12-station criterion.

Discussion

There are a number of standard-setting methods described in the OSCE and performance-based–assessment literature. The currently used modification of the mean-borderline-group method of computing station cut scores in multi-site examinations is the only method that gives unequal weighting to judges (examiners) as a result of the practicalities of implementation. The observed ranges of examiner-specific cut scores, combined with the differences in proportions of borderline candidates identified by different examiners of the same station, open up the potential for individual examiner(s) to influence the cut score in a manner inconsistent with the opinions of the other examiners. This is a sharp contrast with most other standard-setting models, where all experts' judgments are weighted equally. This study proposed an alternative method that attempts to correct for this imbalance and examined the practical implications of using this alternative.

Given the observed variability of examiners in the application of the borderline rating, there was surprisingly little impact when empirical data were subjected to the alternative computation method. There was remarkable consistency between the cut scores generated by the two methods within stations, with the largest observed difference being only 5%, or the equivalent of no more than two checklist items. The small differences translated to similarly consistent pass–fail decisions at the level of individual stations. Of the 78 stations on which the two methods were compared, decisions were unaffected at 58 (74%). Furthermore, neither raw cut scores nor station-level failure rates were systematically affected by use of the alternative method. That is, equal weighting of examiner opinions did not consistently result in more or less stringent cut scores.

The already small effect was further attenuated at the level of pass–fail decisions for the four 19- or 20-station examinations. Concordance rates between the two methods were very high, with final decisions being unaffected for 97% of all candidates included in the analyses. It appears that because there is no systematic effect of the alternative method at the station level, increases in failure rates at one station are being counteracted by decreases at another station in the same examination.

Further, equal weighting of examiner opinions does not influence

30

the reproducibility of station cut scores (and the resultant failure rates) across testing occasions. The degree to which failure rates changed from one use of a given station to the next was not systematically altered by introduction of the new computation method.

It should be noted that this study examined only the effect of the alternative computation method on checklist-based ratings of candidate performance. Station composite scores, which the literature suggests are more reliable,[5] often form the basis for both station-level and exam-level decisions. The station composite scores were not examined in this study due to the complicating effects of measuring multiple constructs with multiple scoring rubrics. The extent to which the findings on checklist scores would be replicated on station composites is still untested but may be of interest to test developers and administrative bodies that use composite scores to assess performances on multidimensional examinations.

The extent to which the observed differences in borderline ratings are related to differences in the candidates' abilities across sites or differences among examiners in their use of the "borderline competent" rating may be of interest to some but is essentially an academic argument. Reasons for observed variations in the frequencies of borderline ratings used by examiners have not been studied to date, and could not be determined through the study design used in this project. Variability in applying the borderline rating was in fact observed in the data set used for the study. For these data, despite differences among examiners, the practical implications of weighting their opinions according to liberality of use of the "borderline" rating do not suggest a need to change current practices in large-scale, multi-site OSCEs. Given the equivocal psychometric benefits of one approach versus the other, a decision about which computation method should be employed when using the mean-borderline-group technique should be based on philosophical and practical rationales.

The author acknowledges the Canadian Alliance of Physiotherapy Regulators for providing data used in the study, and thanks Drs. Arthur Rothman and Glenn Regehr for their helpful comments throughout the study.

Correspondence: Jodi Herold McIlroy, University of Toronto, BW6-668, 585 University Avenue, Toronto, ON M5G 2C4 Canada; e-mail ⟨jodi.herold@utoronto.ca⟩.

References

1. Cizek GJ. Setting passing scores. Educ Meas Issues Pract. 1996;Summer:20–31.
2. Livingston SA, Zieky MJ. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, NJ: Educational Testing Service, 1982.
3. Berk RA. A consumer's guide to setting performance standards on criterion-referenced tests. Rev Educ Res. 1986;6:137–72.
4. Cusimano MD. Standard setting in medical education. Acad Med. 1996;71(10 suppl):S112–20.

31

An Investigation of the Impacts of Different Generalizability Study Designs on Estimates of Variance Components and Generalizability Coefficients

L. A. KELLER, K. M. MAZOR, H. SWAMINATHAN, and M. P. PUGNAIRE

In recent years, performance assessments have become increasingly popular in medical education. While the term "performance assessment" can be applied to many different types of assessments,[1] in medical education this term usually refers to some sort of simulated patient encounter, such as an objective structured clinical examination (OSCE) or a computer simulation of an encounter. These types of assessments appeal to many educators because the tasks or items used are often seen as more realistic than items on multiple-choice examinations. However, this increased "realism" or apparent authenticity comes at a cost—performance examinations are typically more time-consuming and expensive both to administer and to score. On an OSCE, each encounter with a standardized patient is typically scored as a single item, often resulting in an examinee's completing only four to eight items in a two-hour testing period. In contrast, an examinee might complete 100 to 150 items during a two-hour multiple-choice examination.

The fact that performance examinations are typically relatively short means that test users must pay particular attention to the reliability and validity of test scores. In general, other things being equal, a shorter test will result in scores that are less reliable than a longer test. Lower reliability reflects greater error. Adding more items is one way that test developers may increase reliability. On a multiple-choice test, it is relatively inexpensive to write and administer additional items. However, on a performance test both the development and administration of even a single new item can be expensive, and often must be justified in terms of expected gains in score precision.

A second consideration in performance examinations is that scoring is typically more difficult and expensive than scoring of multiple-choice examinations. Expert or trained raters are generally required to review each performance or a sample of performances. Such ratings may be used to score specific performances or to develop scoring criteria or weighting schemes. In either case, raters are a potential source of error.

Generalizability theory[2] provides a framework for estimating the relative magnitudes of various sources of error in a set of scores. In most performance assessments, both items and raters are potential sources of error. Generalizability theory allows estimation of the error associated with each of these sources separately, as well as the relevant interaction effects. In a generalizability study (G study), the variance in a set of scores is partitioned in a manner similar to that used in the analysis of variance. However, in a G study the emphasis is not on testing for statistical significance, but rather on assessing the relative magnitudes of the variance components. Depending on the study design, different variance components can be estimated. Once the variance components are estimated, additional analyses can be conducted. In the framework of generalizability theory, the second stage of analysis is referred to as a decision study (D study). In a D study, the estimated variance components are used to estimate generalizability coefficients (comparable to reliability coefficients) under various measurement conditions. Thus, using the results from a single test administration, it is possible to estimate the impacts of changing both the number of raters and the number of items. This is an important benefit of conducting analyses based on generalizability theory. However, it must be stressed that the variance components and G coefficients are estimates, and as such will vary depending on the specific sample used.

Given that the results of generalizability analyses are often used to make practical decisions about test implementation, it is important to collect the data for a G study in a way that will maximize the precision of the variance-components estimates. Given also that performance assessments are costly to administer and score, and that resources (time, raters, and money) are typically limited, the question of how available resources should be allocated for a G study is an important one. Is it preferable to collect data from 100 examinees on 16 items, or 200 examinees on eight items? Should four raters score 50 examinee performances, or should two raters score 100 performances? Decision studies may help to inform these types of decisions after the data are collected and analyzed, but D studies are based on G studies. To date there is no research we are aware of to help in planning data collection for a G study, especially under constraints.

The purpose of the present study was to examine the impacts of different G-study designs. All of the designs simulated here contain the same number of data points, but the distributions of the data points over examinees, items, and raters are varied. By starting with a relatively large data set (200 medical student examinees, completing 16 items each, scored by four raters each for a total of 12,800 data points), we were able to conduct repeated sampling of different data-collection conditions and to construct empirical confidence intervals for variance components estimates. Computed confidence intervals were also constructed[4] and compared with the empirically constructed intervals. A series of D studies was then conducted to illustrate how different sampling strategies and different samples within those strategies could have substantial impacts on the decisions that would be likely to be made based on such analyses. It should be stressed that the focus of this study was to illustrate the impacts of various sampling strategies, rather than to make decisions about this particular data set. We hope to inform and remind test designers and users that estimates are based on samples, and as such contain variability, and to illustrate the extent to which that variability is greatly affected by the data-collection procedure used.

Method

Data. The data set used here, hereafter referred to as the "full sample," consisted of four expert ratings of 200 medical students on 16 performance items related to a computer simulation. Each examinee performance was rated by each of the four *independent* raters on a holistic nine-point rating scale. From this data set, samples were selected to five data-collection designs or conditions. The numbers of persons or examinees (P), items (I), and raters (R) for each condition were as follows: condition 1, P = 25, I = 16, R = 4; condition 2, P = 50, I = 8, R = 4; condition 3, P = 50, I = 16, R = 2; condition 4, P = 100, I = 4, R = 4; condition 5, P = 100, I = 8, R = 2. These five conditions were chosen so that all samples contained the same total number of observations (1,600). While many other possible combinations were possible, it was beyond the scope of the present study to investigate every possible design. These five conditions were considered representative and realistic. One hundred replications were conducted for each condition in constructing the empirical confidence intervals. For the computed confidence intervals for conditions 1 through 5, one sample was

Table 1. Empirical and Computed 95% Confidence Intervals for the Percentage of Variance Accounted for by Each Source by Condition, University of Massachusetts Medical School, 1999–2000

	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Full Sample
No. of examinees:	25	50	50	100	100	200
No. of items:	16	8	16	4	8	16
No. of raters:	4	4	2	4	2	4
Source						
Person (P)	14.5%	14.2%	14.9%	14.8%	15.1%	14.9%
Empirical	(6.0, 26.9)	(5.4, 22.8)	(8.9, 21.3)	(6.1, 27.2)	(7.4, 23.6)	
Satterthwaite	(8.1, 33.1)	(5.5, 22.7)	(8.7, 24.0)	(9.1, 29.7)	*(24.2, 56.8)*	(11.9, 19.4)
Item (I)	11.9%	13.0%	11.8%	13.3%	12.0%	11.7%
Empirical	(6.7, 17.7)	(.9, 25.1)	(6.8, 16.3)	(.2, 36.2)	(1.2, 25.9)	
Satterthwaite	(6.1, 34.4)	(5.5, 45.1)	(8.7, 43.8)	*(16.1, 39.0)*	(9.4, 74.7)	(5.8, 26.1)
Rater (R)	0.7%	0.7%	0.9%	0.5%	0.7%	0.7%
Empirical	(.3, 1.3)	(0, 2.1)	(−.1, 2.6)	(−.3, 2.3)	(−.4, 3.4)	
Satterthwaite	(.3, 2.6)	*(.8, 5.5)*	(.6, 3.6)	*(−.2, 0)**	*(0, 0)*	(.3, 2.3)
PI	52.2%	52.4%	51.4%	52.2%	51.5%	52.2%
Empirical	(45.5, 60.2)	(41.0, 65.4)	(41.6, 62.7)	(34.9, 71.7)	(37.1, 67.9)	
Satterthwaite	(45.0, 62.0)	(49.1, 67.5)	(41.1, 52.9)	(39.5, 56.2)	*(25.5, 32.3)*	(50.1, 55.9)
PR	.08%	0.1%	0.1%	0%	0%	0.1%
Empirical	(−.4, .4)	(−.4, .7)	(−.3, .6)	(−.8, .8)	(−.6, 1.1)	
Satterthwaite	(0, .3)	*(−.5, 0)**	(0, 1.2)	*(−2.3, −.2)**	*(0, 0)*	(0, .3)
IR	2.4%	2.4%	2.5%	2.3%	2.7%	2.4%
Empirical	(1.5, 3.5)	(.9, 4.3)	(.8, 4.5)	(.6, 5.6)	(.5, 6.2)	
Satterthwaite	(1.4, 4.3)	(1.8, 6.9)	(1.5, 7.5)	(.5, 3.9)	*(.3, 2.3)*	(1.4, 3.5)
Residual	18.1%	17.3%	18.4%	16.9%	18.0%	18.0%
Empirical	(15.7, 20.7)	(14.8, 19.6)	(14.3, 23.3)	(12.9, 21.5)	(13.7, 22.6)	
Satterthwaite	(16.7, 19.9)	*(14.8, 17.7)*	(17.1, 21.0)	*(12.7, 15.2)*	*(13.3, 16.4)*	(17.7, 18.8)

*Due to negative estimates of this variance component, the confidence interval computed using Satterthwaite's technique is not appropriate and should not be interpreted. Italicized confidence intervals do not contain the variance percentage found with the full sample.

selected at random, and computations were based on that single sample.

Analysis. For each of the 500 samples, and for the full data set, a person × item × rater (p × i × r) G study was performed, and variance components were estimated using GENOVA.[3] The 100 replications of each sampling condition provided an empirical sampling distribution for each of the variance components and allowed empirical estimation of means, standard deviations, and 95% confidence intervals for each variance component. The percentage of variance due to each variance component was also calculated, along with the appropriate 95% confidence intervals for these percentages. These empirical confidence intervals were compared with the confidence intervals obtained using Satterthwaite's technique.[4]

To assess the practical implications of the differences in the variance components, a series of D studies was conducted. Because the results of the G studies suggested that only a small percentage of the variance was associated with the rater facet, the number of raters was fixed at four for all D studies, while the number of items varied from one to 30. Two sets of D studies were conducted for each of the five simulated conditions. This was done in order to illustrate how results could differ even under the same data-collection design. The specific samples were chosen so that the person variance component was at the 10th and 90th percentiles of the distribution for that condition. A D study was also conducted on the full data set.

Results

The results of the G studies using the full data set and the five different conditions are summarized in Table 1. For conditions 1 through 5, the percentages associated with each variance component represent the averages across the 100 replications. The confidence intervals reported here are based on the empirical distributions of these percentages in the 100 replications. The confidence intervals obtained using Satterthwaite's technique are reported below the empirical confidence intervals.

Comparing the average percentage of variance associated with each of the facets across the five sampled conditions, it appears that differences between conditions are minimal. The average percentages are also very similar to the variance-components estimates obtained using the full data set. However, because the results for the various sampling conditions were based on the variance components averaged across 100 samples, it is important to consider the associated confidence intervals, which indicate the variability in the sampling distributions. A review of the empirical confidence intervals suggests differences in the stability of the estimates obtained under various conditions. For example, the widths of the empirical confidence intervals for the item component range from about 9% (condition 3) to 36% (condition 4), suggesting that condition 3 provides a more stable estimate of the item-variance component. Considering all five sampling conditions, condition 1 provides the most stable estimates of four of the seven variance components. By contrast, condition 4 provides the least stable estimates of five of the seven components.

The computed confidence intervals show considerable variability across conditions in the widths of the intervals and the values of the lower and upper limits. Sixteen of the 35 computed confidence intervals for conditions 1 through 5 were wider than the empirical intervals; the remaining 19 were not. Twelve of the 35 computed confidence intervals did not contain the value of percentage of

TABLE 2. Estimates of G Coefficients from D studies Based on Variance Components from the 10th and 90th Percentiles of the Five Conditions and the Full Data (No. of Raters = 4), University of Massachusetts Medical School, 1999–2000

	Condition 1		Condition 2		Condition 3		Condition 4		Condition 5		Full Sample
No. of examinees:	25		50		50		100		100		200
No. of items:	16		8		16		4		8		16
No. of raters:	4		4		2		4		2		4
No. of items	10th %	90th %	10th %	90th %	10th %	90th %	10th %	90th %	10th %	90th %	
1	.14	.25	.15	.28	.17	.24	.16	.25	.14	.26	.21
2	.25	.40	.27	.43	.30	.39	.27	.40	.24	.41	.34
3	.33	.50	.35	.54	.39	.49	.36	.50	.33	.51	.44
4	.40	.57	.42	.61	.46	.56	.43	.58	.39	.58	.51
5	.46	.63	.48	.66	.51	.62	.49	.63	.45	.63	.57
6	.50	.67	.52	.70	.56	.66	.53	.67	.49	.68	.61
7	.54	.70	.56	.73	.60	.69	.57	.70	.53	.71	.65
8	.57	.73	.59	.75	.63	.72	.61	.73	.56	.73	.68
9	.60	.75	.62	.77	.66	.74	.64	.75	.59	.76	.70
10	.63	.77	.64	.79	.68	.76	.66	.77	.62	.78	.72
11	.65	.79	.66	**.81**	.70	.78	.68	.79	.64	.79	.74
12	.67	**.81**	.68	.82	.72	**.80**	.70	**.80**	.66	**.80**	.76
13	.69	.81	.70	.83	.73	.81	.72	.81	.68	.82	.77
14	.70	.82	.72	.84	.75	.82	.74	.83	.69	.83	.79
15	.72	.83	.73	.85	.76	.83	.75	.83	.71	.84	**.80**
16	.73	.84	.74	.86	.77	.84	.76	.84	.72	.84	.81
17	.74	.85	.75	.87	.78	.85	.78	.85	.73	.85	.82
18	.75	.86	.76	.87	.79	.85	.79	.86	.75	.86	.82
19	.76	.86	.77	.88	**.80**	.86	**.80**	.86	.76	.87	.83
20	.77	.87	.78	.88	.81	.87	.81	.87	.77	.87	.84
21	.78	.88	.79	.89	.82	.87	.81	.88	.77	.88	.85
22	.79	.88	**.80**	.89	.82	.88	.82	.88	.78	.88	.85
23	.80	.89	.80	.90	.83	.88	.83	.89	.79	.89	.86
24	**.80**	.89	.81	.90	.83	.89	.84	.89	**.80**	.89	.86
25	.81	.89	.82	.90	.84	.89	.84	.89	.80	.89	.87

Note: Numbers of items estimated to be needed to obtain a G coefficient of .80 are shown in bold.

variance estimated from the full sample, and 12 did not contain the value of the mean percentage of variance estimated from the 100 samples of the specified condition.

As noted above, a series of D studies was conducted to illustrate how estimates of G coefficients might vary depending on the sampling design and the specific sample used in the G study. Because such decision studies are often used to determine a minimal number of items to be administered to obtain a specified G coefficient (much as the Spearman–Brown prophecy formula is used in classical test theory), the number of items was varied from 1 to 25. These results are presented in Table 2. One result of interest is the number of items estimated to be needed to obtain a G coefficient of .80. This value is in bold in each column.

Considering the 90th percentile samples for all five conditions, it can be seen that for four of the five conditions the estimate of the number of items needed to achieve a G of .80 is 12; for the second condition the estimate is 11. The estimate based on the full sample is 15 items. Considering the 10th percentile samples for the five sampling conditions, the estimated numbers of items needed range from 19 to 24. Comparing the 10th and 90th percentile samples within conditions, substantial differences in estimates are apparent. For instance, for condition 5 (N_p = 100, N_i = 8, N_r = 2) the number of items estimated to be necessary at the 10th percentile is 24, versus only 12 items if the sample at the 90th percentile is used.

Discussion and Implications

The results presented above suggest that, at least for the data set used here, different data-collection designs would have had little impact on average on the variance-components estimates obtained. In other words, collecting ratings of 25 examinees on 16 items using four raters would have resulted in approximately the same variance-components estimates as collecting ratings of 100 examinees on eight items using two raters. In all the conditions studied here, including the full sample, it was clear that a far higher percentage of the variability in scores was related to the item facet, and the associated interactions, compared with that associated with the rater facet and associated interactions. Thus, conclusions as to the relative impacts of item and rater would have been similar regardless of which data-collection design was used.

While the average percentages of variance associated with the individual facets were very similar across the five conditions, the widths of the associated confidence intervals (both empirical and computed) did vary across conditions. In addition, for all of the five conditions, estimates of the numbers of items needed to obtain a given level of generalizability varied considerably depending on whether the 10th percentile or the 90th percentile sample was used. Since in practice investigators have only one sample, and no knowledge of where their sample falls in the distribution, it is important to be aware of the fact that substantially different estimates might have resulted if a different sample had been used. The results of the D studies reported here highlight this. Depending on the condition, the numbers of items required to obtain a G of .80 differed by as much as 100% depending on the specific sample used in the G study. This is particularly important given the time and expense associated with most performance assessments. In this case, analysis of the full data set suggests that 15 items would be needed

to achieve a G coefficient of .80. While this is also, of course, a sample, it can be considered our best estimate of the true number of items needed. If, instead of the full data set, we had had only one of the samples investigated here, we might have come to the conclusion that a test of only 11 or 12 items would result in a G coefficient of .80. Were we then to administer a 12-item test based on these results, it is likely that the results would be less generalizable than expected, a result with potentially serious consequences for test developers and users. In contrast, using a different sample we might conclude that 24 items were needed to obtain sufficiently generalizable results. While this overestimation would not be a problem from a psychometric perspective, the costs associated with administering more items than are in fact needed to achieve a specified G could be. In fact, in some circumstances, an overestimation error could result in a decision that a particular testing format is not feasible given cost and time constraints.

It is difficult to interpret the results found for the computed confidence intervals, particularly since these were calculated based on a single sample and would be expected to differ if a different sample were selected. In comparing the computed confidence intervals with the empirical confidence intervals, substantial discrepancies were found, especially for the components that accounted for higher percentages of the variance. In some cases discrepancies were found not only in the width of the interval but also between the values within the interval—at times the intervals were not even overlapping. These findings raise questions as to the usefulness of Satterthwaite's technique in this instance.

Conclusions

It is important for test developers, psychometricians, and test users to remember that generalizability coefficients and other reliability coefficients are estimates based on samples, and as such may be expected to vary depending on the specific sample used in estimation. The results of the study reported here highlight how different samples may produce different results, which can in turn lead to very different decisions. The fact that the computed confidence intervals differed substantially from the empirical confidence intervals, and in some cases did not even contain the appropriate percentage, suggests that computing confidence intervals from a single sample will not necessarily improve decision making. This study does not allow us to make specific recommendations regarding the number or distribution of data points required when conducting a G study. Which design provides the most stable estimates will depend on the nature of the data collected. It is ideal, naturally, to obtain the largest sample possible; but when smaller samples are used, it is crucial that the stability of the estimate be taken in to consideration before decisions are made based on a specific sample, as was illustrated in this study.

Correspondence: Lisa Keller, 152 Hills South, University of Massachusetts, Amherst, MA 01003; e-mail ⟨lisa@pumpingstation.com⟩.

References

1. Kane M, Crooks T, Cohen A. Validating measures of performance. Educ Meas Issues Pract. 1999;18:5–14.
2. Brennan RL. Elements of Generalizability Theory. Iowa City, IA: ACT Publications, 1992.
3. Crick JE, Brennan, RL. Manual for GENOVA: A Generalized Analysis of Variance System. Iowa City, IA: ACT Publications, 1983.
4. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics Bull. 1946;2:110–4.

35

● **CLOSE BUT NO BANANAS: PREDICTING PERFORMANCE**

Moderator: Mark Albanese, PhD

A Validity Study of the Writing Sample Section of the Medical College Admission Test

MOHAMMADREZA HOJAT, JAMES B. ERDMANN, J. JON VELOSKI, THOMAS J. NASCA, CLARA A. CALLAHAN, ELLEN JULIAN, and JEREMY PECK

The current version of the Medical College Admission Test (MCAT), introduced in 1991, includes four sections: Biological Sciences, Physical Sciences, Verbal Reasoning, and Writing Sample. The Writing Sample assesses skills in organizing thoughts and presenting ideas in a cohesive manner, and provides evidence of analytic thinking and writing skills.[1] Scoring is based on two 30-minute essays about general topics. An example of an essay prompt is "In a free society, individuals must be allowed to do as they choose."

Each essay is holistically scored by two trained reviewers on a six-point scale with regard to specific criteria such as developing the central idea, synthesizing concepts logically, and writing clearly with good grammar, syntax, and punctuation. Essays receiving scores that differ by more than one point are re-evaluated by a third expert reviewer. The scores for the two essays completed by each examinee are summed and converted to an 11-point alphabetical scale ranging from J to T. According to reports by the Association of American Medical Colleges (AAMC), 98% of the essays are given identical scores or scores within one scale point of each other by the independent reviewers.[1]

The results of multi-institutional studies, conducted by the MCAT Validity Study Advisory Group,[2] have been published and presented at professional meetings.[2–5] However, while the need for additional studies of the psychometric properties of the MCAT continues, there is a particular need for study of the predictive power of the Writing Sample. The unique alphabetic scores of the Writing Sample discourage the usual correlational analyses used in validity studies. Although it is possible to convert the alphabetic scores to the integers from 1 to 11 by assuming that the letters constitute an interval scale, such an assumption might not be widely accepted.

We designed the present study to examine the validity of the Writing Sample section of the MCAT for students at Jefferson Medical College in Philadelphia, Pennsylvania. We speculated that the ability to organize and express ideas effectively in writing could have relevance to the analytic and problem-solving skills demanded in clinical performance. Furthermore, such skills might also be related to a better presentation of one's self, and to effective verbal expression of ideas, both of which are critical in promoting interpersonal relationships. Therefore, we hypothesized that scores on the Writing Sample would be associated more closely with indicators of clinical competence than with measures of achievement in basic sciences.

Method

Data for 1,776 matriculants (1,086 men, 690 women) at Jefferson Medical College between 1992 and 1999 were retrieved from the database of the Jefferson Longitudinal Study of Medical Education.[6] The students were classified into three groups (top, middle, and bottom) based on their scores on the Writing Sample. The "top" group included 314 (18% of the sample) who scored R, or S, or T. The "middle" group consisted of 1,115 (65%) who scored N, O, P, or Q. The 307 (17%) students who scored J, K, L, or M comprised the "bottom" group.

Three sets of criteria were used.

- *Admission measures*. The first set included the measures typically used for screening applicants, such as undergraduate grade-point averages (UGPAs) in science and non-science courses, admission interview scores, and MCAT scores on Biological Sciences, Physical Sciences, and Verbal Reasoning.
- *Performance in the basic sciences*. The second set consisted of achievement measures in the basic science disciplines, including grade-point averages (GPAs) in first- and second-year medical school courses. Scores on Step 1 of the United States Medical Licensing Examinations (USMLE) were also included.
- *Performance in clinical sciences and ratings of clinical competence*. Included in this set were scores on written examinations in six core clerkships (family medicine, internal medicine, obstetrics–gynecology, pediatrics, psychiatry, and surgery) in the third year of medical school. Written examinations in basic and clinical sciences are in either multiple-choice or uncued formats,[7] with reliability estimates usually over $r = .75$.

Combined global ratings of clinical competence in the six core clerkships, on a 100-point scale,[8] and scores on Step 2 of the USMLE were also included. In addition, medical school class rank (percentile), a composite measure with two thirds weight for clinical competence in the core clerkships and one third weight for the combined first- and second-year GPAs,[8,9] was used, as were the ratings of graduates' clinical competence from a 33-item rating form measuring three clinical competence areas of "data-gathering and processing skills" (16 items), "interpersonal skills and attitudes" (ten items), and "socioeconomic aspects of patient care" (seven items). These ratings were made on a four-point Likert scale by program directors near the end of the first postgraduate year. Data have been reported in support of the measurement properties of this rating form, including construct validity (factor structure), the internal consistency aspect of reliability, and criterion-related validity.[10,11]

Continuous measures were transformed to a distribution with a mean of 100 and a standard deviation of 10 to facilitate comparisons of the magnitudes of differences on a scale with a uniform mean and standard deviation. This transformation was used to mitigate the issue of scale incompatibility within each class and between classes. The numbers of observations vary for different analyses because data were not yet available for the entire sample at the time of this study.

The three groups were compared with respect to the criterion measures by using analysis of variance for continuous measures, followed by the Duncan test and the Kruskal–Wallis test for class rank. Analysis of covariance was also employed to make statistical adjustments for baseline differences in the scores of other MCAT sections.

Results

Admission Variables. The means and sample sizes for the criterion measures and a summary of the statistical analyses are presented in Table 1. Comparisons of the top, middle, and bottom groups on the Writing Sample showed no significant difference for undergraduate science GPA, or for the Biological and Physical Sciences

36

Table 1. Means of Selected Admission Measures and Performances in Basic and Clinical Sciences by Scores on the Writing Sample Section of the Medical College Admission Test (MCATP)*

Criterion Measure‡	Mean Criterion Measures for Three Groups Classified by Level of Writing Sample Scores† Top (n = 314)	Middle (n = 1,155)	Bottom (n = 307)	Effective n§	F-ratio	p
Admission						
Undergraduate GPAs: science	100.2	100.0	99.7	1,766	.21	.81
Undergraduate GPAs: non-science	101.3	99.9	99.1	1,766	4.3	.02
Admission interview	100.2	100.0	99.6	1,769	.30	.74
MCAT: Biological Sciences	100.2	99.6	99.0	1,776	1.3	.28
MCAT: Physical Sciences	100.2	99.4	99.1	1,776	1.2	.28
MCAT: Verbal Reasoning	103.1	99.5	96.8	1,776	31.9	<.01
Basic sciences						
Medical school: 1st- & 2nd-year GPAs	100.9	100.0	99.2	1,535	1.8	.17
USMLE: Step 1	100.8	99.9	99.7	1,271	.97	.38
Clinical sciences and ratings of clinical competence						
Medical school: 3rd year (objective tests)	101.9	100.0	98.0	1,036	6.0	<.01
Medical school: 3rd year (clinical ratings)	101.9	100.0	98.0	1,036	5.9	<.01
Medical school class rank	86.0	85.5	84.8	1,036	6.1	<.01
USMLE: Step 2	100.6	99.9	98.7	1,006	3.1	.04
Postgraduate ratings: data gathering	103.1	100.0	97.9	433	2.7	.07
Postgraduate ratings: interpersonal & attitudes	102.8	100.0	97.1	433	3.2	.04
Postgraduate ratings: socioeconomics of patient care	102.8	100.0	97.7	433	2.5	.08
Postgraduate ratings: physician as a clinician	103.2	99.6	98.5	423	2.4	.09
Postgraduate ratings: physician as an educator	103.4	99.0	97.8	364	2.8	.06
Postgraduate ratings: physician as a manager	101.5	99.5	98.4	339	.71	.49

* Participants were 1,776 students who entered Jefferson Medical College between 1992 and 1999.

† "Top" category includes R, S, T; "middle" category includes N, O, P, Q; and "bottom" category includes J, K, L, or M alphabetic scores.

‡ With the exception of medical school class rank, all other criterion measures for each entering class were transformed to a distribution with a uniform mean of 100 and a standard deviation of 10 to facilitate comparisons of mean differences.

§ Numbers of observations vary due to unavailability of data at the time of the study.

sections of the MCAT. However, significant differences were observed for undergraduate non-science GPA ($p < .05$), and the Verbal Reasoning test ($p < .01$). Duncan tests indicated that the top group's undergraduate non-science GPA was significantly higher than those of the middle and bottom groups ($p < .05$). As expected, the top group also obtained the highest mean score in Verbal Reasoning, followed by the middle and bottom groups ($p < .01$).

Performances in Basic Sciences Disciplines in Medical School. Data reported in Table 1 indicate that although the top group consistently outperformed the bottom group in first- and second-year basic science courses, as well as on USMLE Step 1, the differences were not statistically significant.

Performances in Clinical Science Disciplines and Ratings of Clinical Competence. Statistically significant differences were observed among the top, middle, and bottom groups on a number of performance measures in clinical disciplines. Both the top and the middle groups obtained significantly higher mean grades ($p < .01$) than did the low group on written examinations in the six core clerkships. A similar pattern of findings was observed for medical school class rank.

The top group was also rated significantly higher than the middle and bottom groups in global ratings of clinical competence in the third-year core clerkships ($p < .01$). The difference between the top and bottom groups' Step 2 scores was also statistically significant ($p < .05$).

Results for the six measures of clinical competence in residency showed that the differences for ratings in interpersonal skills and attitudes were statistically significant ($p < .05$), where the top group was rated significantly higher than the bottom group. Although the differences in other areas of postgraduate competence did not reach the conventional level of statistical significance ($p < .05$), a consistent pattern was observed in which the highest average ratings were obtained by the top group, and the lowest by the bottom group.

In additional analyses, the two extreme groups (top and bottom) were compared regarding the ratings in other areas of clinical competence in residency, and standardized effect-size estimates (d) were calculated for the significant pairwise differences. The top group was rated higher than the bottom group in data-gathering and processing skills ($p < .05$, estimated effect size = .52), socioeconomic aspects of patient care ($p < .05$, effect size = .51), and physician as a patient educator ($p < .05$, effect size = .56). Effect-size estimates of this order of magnitude are not small according to Cohen's definition.[12] These differences are not only statistically significant, but also of practical significance.

Controlling for Differences on the Other Sections of the MCAT. Statistical adjustments were made for baseline differences using both the Biological Sciences and the Physical Sciences sections of the MCAT as covariates through analysis of covariance. Each of the previously-reported differences among the three groups remained unchanged. This confirms that the previous findings were not confounded by score differences in these two sections of the MCAT.

Further statistical adjustments were made by adding scores on the Verbal Reasoning section of the MCAT to the other two covariates (scores on the Biological and Physical Sciences sections). The differences remained unchanged for the following criterion measures: clinical clerkship examinations (adjusted p = .02), clinical clerkship ratings (adjusted p = .02), and medical school class rank (adjusted p = .008). However, changes in statistical significance levels were observed in the undergraduate non-science GPAs (adjusted p = .10), Step 2 scores (adjusted p = .31), and postgrad-

uate ratings of data-gathering and data-processing skills (adjusted $p = .11$).

Discussion

The findings of the present study confirm the research hypothesis that scores on the Writing Section of the MCAT yield a closer association with measures of clinical competence than with achievement in the basic sciences.

These findings provide support for the validity of the Writing Sample from a number of perspectives. We hypothesized that high scorers on the Writing Sample would outperform others in clinical sciences evaluations and in ratings of clinical competence. The hypothesis was confirmed, providing support for the predictive validity of the test.

The fact that scores on the Writing Sample were significantly associated with performance in the clinical areas in medical school and residency provides evidence in support of convergent validity, whereas their lack of associations with measures of achievement in science prior to and during medical school supports the discriminant validity of the test. In addition, concurrent validity was demonstrated by the relationships between the Writing Sample and Verbal Reasoning scores.

Clinical grades in medical school are based on the observations of faculty and supervising residents during the actual provision of clinical care to patients, and reflect the ability of students to relate well to others. These dimensions of clinical competence require basic medical knowledge, which may be predicted on the basis of MCAT science scores. However, while necessary, medical knowledge is not sufficient for effective clinical decision making. The significant relationship between the Writing Sample scores and clinical ratings after adjustment for MCAT science scores confirms that the associations between Writing Sample scores and measures of clinical performance are beyond those that would be expected from attainment of knowledge only. Therefore, it can be concluded that the Writing Sample measures a unique skill, different from those measured by the other sections of the MCAT, including the Verbal Reasoning section. It can be speculated that such a unique skill might be attributed more to factors that are not associated with achievement in sciences. Such speculation needs to be verified further by empirical evidence.

The results generally suggest that, for a sample of students at one medical school, Writing Sample scores of J, K, L, or M predicted poorer clinical performance during and after medical school. This particular grouping of the Writing Sample scores should be studied further in samples from other medical schools before implementation in decision making.

Certain aspects of this study could be questioned and deserve comment. It may be argued that the statistically significant findings of this study could have been due to chance as a result of the large number of statistical comparisons that were performed. However, this argument can be refuted based on the findings for the 18 criterion measures reported in Table 1. While only one statistically significant finding would be expected by chance alone at $p < .05$, seven were reported in this table. Similarly, the internal validity of the findings could be questioned by arguing that the statistically significant findings could be attributed to the large sample size, rather than underlying relationships among the variables. This argument can also be refuted based on the findings that the significant associations were observed only for the conceptually relevant scores, such as Verbal Reasoning, whereas there was no relationship with the less relevant scores such as the Biological and Physical Sciences, despite the fact that the sample size was equally large ($n = 1,776$) in all analyses. Furthermore, the magnitudes of the effect-size estimates between top and bottom scorers suggest that the obtained differences are of practical importance to decision makers.

These findings, coupled with the relatively large sample size and the longitudinal design of this study, provide assurance for the internal validity of the results. However, more data from other medical schools are needed to assure the external validity and the generalization of the findings.

In earlier studies we found that validity coefficients for the MCAT varied for students who graduated from different colleges and universities,[13] that the validity of the MCAT varied for different sets of scores when applicants repeated the examination,[14] and that different sections of the MCAT have different predictive validity depending upon the criterion measures.[15] Empirical evidence also suggests that validity coefficients for the MCAT vary among medical schools.[2] It will be essential to consider these factors in future studies of the validity of MCAT.

Correspondence: Mohammadreza Hojat, PhD, Jefferson Medical College, Philadelphia, PA 19107; e-mail: (*Mohammadreza.Hojat@mail.tju.edu*).

References

1. Association of American Medical Colleges. Use of MCAT Data in Admissions: A Guide for Medical School Admissions Officers and Faculty. Washington, DC: AAMC, 1991.
2. Koenig JA, Wiley A. Medical school admission testing. In: Dillon RF (ed). Handbook of Testing. West Port, CT: Greenwood Press, 1997:274–95.
3. Mitchell K, Haynes R, Koenig JA. Assessing the validity of the updated Medical College Admission Test. Acad Med. 1994;69:394–401.
4. Wiley A, Koenig JA. The validity of the Medical College Admission Test for predicting performance in the first two years of medical school. Acad Med. 1996; 71(10 suppl):S83–S85.
5. Koenig JA, Sireci SG, Wiley A. Evaluating the predictive validity of MCAT scores across diverse applicant groups. Acad Med. 1998:73:1095–106.
6. Hojat M, Gonnella JS, Veloski JJ, Erdmann JB. Jefferson Medical College's longitudinal study: a prototype of assessment of changes. Education for Health. 1996; 9:99–113.
7. Veloski JJ, Rabinowitz HK, Robeson MR, Young PR. Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. Acad Med. 1999;74:539–46.
8. Blacklow RS, Goepp CE, Hojat M. Class ranking models for dean's letters and their psychometric evaluation. Acad Med. 1991;66(9 suppl):S10–S12.
9. Blacklow RS, Goepp CE, Hojat M. Further psychometric evaluations of a class ranking model as a potential predictor of graduates' clinical competence in the first year of residency. Acad Med. 1993;68:295–7.
10. Hojat M, Veloski JJ. Borenstein BD. Components of clinical competence ratings: an empirical approach. Educ Psychol Meas. 1986;46:761–9.
11. Hojat M, Borenstein BD, Veloski JJ. Cognitive and noncognitive factors in predicting the clinical performance of medical school graduates. J Med Educ. 1988; 63:323–5.
12. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence, Erlbaum, 1987.
13. Zeleznik C, Hojat M, Veloski JJ. Predictive validity of the MCAT as a function of undergraduate institutions. J Med Educ. 1987;62:163–9.
14. Hojat M, Veloski JJ, Zeleznik C. Predictive validity of the MCAT for students with two sets of scores. J Med Educ 1985;60:911–8.
15. Glaser K, Hojat M, Veloski JJ, Blacklow RS, Goepp CE. Science, verbal, or quantitative skills: which is the most important predictor of physician competence? Educ Psychol Meas. 1992;52:395–405.

38

• CLOSE BUT NO BANANAS: PREDICTING PERFORMANCE

Moderator: Mark Albanese, PhD

Prediction of Students' Performances on Licensing Examinations Using Age, Race, Sex, Undergraduate GPAs, and MCAT Scores

J. JON VELOSKI, CLARA A. CALLAHAN, GANG XU, MOHAMMADREZA HOJAT, and DAVID B. NASH

The annual selection of new students is one of the most important activities of medical school faculty. They face the challenge of selecting those who can perform well not only in the preclinical years, but also in the clinical arena of medical school, in graduate medical education, and beyond.[1] To make sound, evidence-based decisions, faculty involved in the admission process depend on empirical studies that examine the relationship of an applicant's academic performance before medical school to that individual's academic performance during medical school and afterwards.

Studies have consistently shown that Medical College Admission Test (MCAT) scores and undergraduate grade-point averages (GPAs) are the most important indicators of students' future academic performances.

Specifically, MCAT science scores and undergraduate science GPAs have been associated with preclinical academic performance.[2] However, verbal scores on the MCAT and non-science GPAs have been more closely associated with performance in the clinical years, such as on the United States Medical Licensing Examination (USMLE) Steps 2 and 3.[3] Correspondingly, the combination of GPAs and MCAT scores has been shown to be the best predictor of preclinical academic performance.[4]

The predictive strength of MCAT scores and GPAs is less clear when students' race and sex have been considered, and when performance has been followed longitudinally beyond the preclinical years. Men on average have outperformed women on the USMLE Step 1. The differences were moderated, but not eliminated, by statistical control for differences in prematriculation measures. Conversely, women have outperformed men on the National Board of Medical Examiners (NBME) Part II, though the differences were not as great as those observed between the scores of men and women on Part I, where men outperformed women.[5] Control for differences in prematriculation measures and Part I performances increased the magnitude of differences between women and men on Part II. This phenomenon had been noted several decades earlier.[6–8] Finally, the findings related to students' ages have been equivocal, often because age has been confounded with sex or undergraduate academic performance.[9]

Studies among racial groups have revealed substantial differences in performances on Part I. Although white students on average have scored highest, followed by Asian Americans, Hispanics, and African Americans, these gaps become narrower after controlling for MCAT scores and undergraduate GPAs.[10] One might expect Asian Americans, who as a group have had the highest mean MCAT scores, to outperform other racial groups during medical school. However, two major studies across time and across medical schools have reported lower mean performance for Asian Americans than for white students in medical school.[10,11]

In summary, previous admission-prediction studies have looked at the predictive value of MCAT scores and GPAs for USMLE Step 1 performances among racial groups,[10,12,13] clerkship performance during medical school,[14] and a combination of Step 1 and clerkship performances.[15] Other studies have ignored either students' age, race, or sex when examining the correlation between prematriculation measures and students' performances during medical school,[16] or have studied characteristics such as race without controlling for GPAs and MCAT scores.[11]

We designed the present study to evaluate simultaneously the relative importances of MCAT scores, undergraduate GPAs, age, race, and sex in predicting performances on the three-step sequence of preclinical, clinical, and postgraduate licensing examinations.

Method

The sample consisted of 6,239 matriculants who entered Jefferson Medical College during the 30 years between 1968 and 1997, inclusive. The dependent variables were total scores on Parts I, II, and III of the licensing examinations of the NBME and total scores on Steps 1, 2, and 3 of the USMLE (the latter three examinations replaced the former three several years ago). Although scores on either of the preclinical examinations (Part I, Step 1) were available for every individual studied, scores on the second, clinical tests (Part II, Step 2) were available for only 5,887, because the others had either left medical school or not yet taken the test. Scores on the Part III and Step 3 examinations were collected prospectively for the 3,884 graduates (62%) who had given written permission and completed the examination at the time of the study.

A separate multivariate linear regression model was generated for each of the six dependent variables. NBME scores were transformed from a mean and standard deviation of 500 and 100 to the USMLE scale of 200 and 20 to allow comparisons across the two time periods. Repeated scores were averaged. MCAT scores in earlier time periods were transformed to the current scale and repeated scores averaged using methods reported previously.[17] Scores on science subtests were averaged to estimate an overall science score. Sex was coded 0 for men and 1 for women, who were 26% of the entire cohort. Students who were more than 23 years old at the time of matriculation (also 26% of the cohort) were coded 1 and others were coded 0. An earlier study of a portion of the cohort confirmed that this age cut-off serves as a proxy for nontraditional students.[9] Racial–ethnic backgrounds, as defined by the Association of American Medical Colleges, consisted of the Asian, Oriental, or Pacific Islander groups (the Asian American group in this study); Hispanic (not white); black; and white. Students in each of the first three race categories were coded as either 1 for those in the group, or 0 for those not. The percentages for Asian American, Hispanic, and black were 8.2%, 1.4%, and 2.8%, respectively. The other students, who included 85.9% white and 1.7% in other racial groups with very small sample sizes, were not coded separately.

Results

Each of the six linear regression models shown in Table 1 was statistically significant (F test, $p < .05$). The proportions of variance explained ranged from a high for Step 1 of .26 to a low for Part III of .15. We report only the linear regression weights for the independent variables that were significant (t-test, $p < .05$). The b-coefficients for independent variables provide information about the absolute contribution of each variable as a predictor of the dependent variable. The beta-coefficients, which are transformations of the b-coefficients to a uniform scale across all independent variables, enable comparisons of the relative importance among the independent variables. For example, the b-coefficient of 4.26 for

39

Table 1. B- and Beta-coefficients, Sample Sizes, and Proportions of Variance for Regressions Predicting Performances on NBME and USMLE Examinations from Applicant Data for Matriculants from 1968 through 1997*

	NBME						USMLE					
	Part I		Part II		Part III		Step 1		Step 2		Step 3	
Variable	B	Beta	B	Beta	B	Beta	B	Beta	B	Beta	B	Beta
MCAT Science	3.13	0.31	2.56	0.23	1.96	0.16	4.26	0.34	3.19	0.22	1.31	0.11
MCAT Verbal	1.32	0.14	2.38	0.21	1.73	0.14	1.49	0.13	2.76	0.21	3.07	0.28
Science GPA	7.40	0.18	7.70	0.18	6.99	0.18	9.41	0.21	9.22	0.18	4.48	0.11
Non-science GPA												
Asian American	−5.52	−0.06	−7.98	−0.09	−9.80	−0.09	−7.60	−0.16	−11.47	−0.21	−7.60	−0.18
African American			−4.00	−0.04	−8.55	−0.05						
Hispanic					−6.60	−0.03						
Woman	−3.63	−0.09							3.54	0.08	3.10	0.09
Older												
Constant	139.48		133.39		142.61		120.67		112.37		146.31	
n	4,299		4,227		3,234		1,940		1,660		650	
R2	0.23		0.21		0.15		0.26		0.23		0.17	

*NBME scores, which were rescaled to a mean of 200 and standard deviation of 20, were available for students who entered before 1989. USMLE scores were used thereafter. Only b-coefficients and beta-coefficients that were significant at $p < .05$ by a t-test that b = 0 are reported. Blank values were not significant.

the MCAT science score in the USMLE Step 1 model indicates that a one-point increment in a student's MCAT science score raises his or her predicted Step 1 score by 4.26 points. Comparison of the beta-coefficient of .34 for the MCAT science score with the beta-coefficient of .21 for science GPA indicates that the unique contribution of the MCAT science score as a predictor of Step 1 is about one and a half times that of the science GPA. The validity of these interpretations of beta-coefficients assumes equivalent variability across independent variables, which has been documented in other published studies of portions of this cohort.[18]

As would be expected, the contribution of the MCAT science score in predicting scores on the preclinical examination was more important than that of the science GPA. Being an older, nontraditional student at matriculation was unrelated to all scores after controlling for the other independent variables. The regression coefficients for women were negative for the NBME Part I, but insignificant for Step 1. However, being a woman was positively associated with the scores on USMLE Steps 2 and 3. Although being black was negatively associated with performances on Parts II and III, and being Hispanic negatively associated with performance on Part III, these patterns disappeared in the more recent USMLE examinations. Overall, the only consistent pattern related to age, race, or sex across all examinations was the negative regression weight for Asian American students.

Discussion and Conclusion

This longitudinal study examined the absolute and relative contributions of MCAT scores and undergraduate GPAs, along with age, race, and sex, in predicting students' performances on the sequence of three licensing examinations over the past three decades. The analysis reflected a large number of subjects, including a small fraction of students who reached Part I or Step 1 but did not graduate from medical school. Although the dependent variables were limited to licensing examinations, these are uniform measures that apply across all medical schools.

As expected from many earlier studies, MCAT scores were consistently more valuable than were undergraduate GPAs as predictors of performance on licensing examinations, supporting their continued use in selection decisions.[19] These relationships are stable across three decades and apply to the three examinations. Verbal scores tended to be better indicators of performances in the clinical and postgraduate tests. Although the non-science GPA never appeared in the six regression models, this may be due to the high correlation (r = 0.61) between science and non-science GPAs. There was no independent effect for older, nontraditional students after controlling for their undergraduate academic performances and MCAT scores.

Earlier studies have indicated that, although underrepresented minorities have entered medical school with significant educational disadvantages and have continued to score lower than other students on some measures, their clinical performances were nearly equivalent to those of other students.[20] In the present study, statistical control of the baseline differences at matriculation using regression analysis showed that underrepresented-minority students compared with white students performed less well than would have been predicted on the NBME in the earlier time period. However, this pattern disappeared in the recent time period. This change over time may have been due to the effectiveness of academic enrichment programs.[21,22] It has been reported that the gap in MCAT scores and undergraduate GPAs between underrepresented minorities on one side and white and Asian American students on the other persists, supporting the need for programs aimed at enhancing students' academic preparation before medical school.[23]

The most striking finding is the large negative value of the b-coefficients as well as the beta-coefficients for Asian American students. This indicates that, after controlling statistically for science and verbal MCAT scores and undergraduate GPAs, these students performed less well compared with white students. Previous studies had revealed that Asian American students' performances during medical school were not as good as those of white students, without controlling for prematriculation measures.[11] However, the differences between Asian American and the underrepresented-minority groups in Step 1 performances were significantly reduced after controlling for prematriculation measures.[10] The findings of the present study indicate that Asian American students' performances fell below expectations on all NBME and USMLE examinations, after controlling for these prematriculation measures.

One possible explanation for the decline in performance from the admission test to the licensing examinations may be that Asian American families are less able to influence academic achievement as their young adults mature. It has been reported that certain Asian American families place greater emphasis on doing well in school than do other groups.[24] However, it is very important to

consider that the sample used in the present study included a heterogeneous mix of Asian American students from families that left diverse cultures in Asia at different points in time. Further studies are needed to evaluate these subgroups, to investigate other measures of academic and clinical performance, and to better understand the factors that may influence Asian American students' performances in medical school and beyond.

Correspondence: Jon Veloski, MS, Center for Research in Medical Education and Health Care, Jefferson Medical College, 1025 Walnut Street, Room 119, Philadelphia, PA 19107; e-mail: ⟨jon.veloski@mail.tju.edu⟩.

References

1. Elam CL, Wilson JF, Johnson R, Wiggs JS, Goodman N. Challenging the system: admission issues for at-risk students, admission committees. Acad Med. 1999;74(10 suppl):S58–S61.
2. Jones RF, Thomae-Forgues M. Validity of the MCAT in predicting performance in the first two years of medical school. J Med Educ. 1984;59:455–64.
3. Glaser K, Hojat M, Veloski JJ, Blacklow RS, Goepp CE. Science, verbal, or quantitative skills: which is the most important predictor of physician competence? Educ Psychol Meas. 1992;52:395–406.
4. Mitchell K, Haynes R, Koenig JA. Assessing the validity of the updated Medical College Admission Test. Acad Med. 1994;69:394–401.
5. Case SM, Becker DF, Swanson DB. Performances of men and women on NBME Part I and Part II: the more things change. Acad Med. 1993;68(10 suppl):S25–S27.
6. Weinberg E, Rooney JF. The academic performance of women students in medical school. J Med Educ. 1973;48:240–7.
7. Willoughby TL, Calkins V, Arnold L. Different predictors of examination performance for male and female medical students. JAMWA. 1979;34:316–20.
8. Arnold L, Willoughby TL, Calkins V, Jensen T. The achievement of men and women in medical school. J Am Med Women's Assoc. 1981;36:213–21.
9. Herman MW, Veloski JJ. Premedical training, personal characteristics and performance in medical school. Med Educ. 1981;15:363–7.
10. Dawson B, Iwamoto CK, Ross LP, Nungester RJ, Swanson DB, Volle RL. Performance on the National Board of Medical Examiners Part I examination by men and women of different race and ethnicity. JAMA. 1994;272:674–9.
11. Xu G, Veloski JJ, Hojat M, Gonnella JS, Bacharach B. Longitudinal comparison of the academic performances of Asian-American and white medical students. Acad Med. 1993;68:82–6.
12. Koenig JA, Sireci SG. Evaluating the predictive validity of MCAT scores across diverse applicant groups. Acad Med. 1998;73:1095–106.
13. Vancouver JB, Reinhart MA, Solomon DJ, Haff JJ. Testing for validity and bias in the use of GPA and the MCAT in the selection of medical school students. Acad Med. 1990;65:694–7.
14. Huff KL, Koenig JA, Treptau MM, Sireci SG. Validity of MCAT scores for predicting clerkship performance of medical students grouped by sex and ethnicity. Acad Med. 1999;74(10 suppl):S41–S44.
15. Silver B, Hodgson CS. Evaluating GPAs and MCAT scores as predictors of NBME I and clerkship performances based on students' data from one undergraduate institution. Acad Med. 1997;72:394–6.
16. Hall FR, Bailey BA. Correlating students' undergraduate science GPAs, their MCAT scores, and the academic caliber of their undergraduate colleges with their first-year academic performances across five classes at Dartmouth Medical School. Acad Med. 1992;67:121–3.
17. Hojat M, Veloski JJ, Zeleznik C. Predictive validity of the MCAT for students with two sets of scores. J Med Educ. 1985;60:911–8.
18. Callahan C, Veloski JJ, Xu G, Hojat M, Zeleznik C, Gonnella JS. The Jefferson–Penn State B.S.–M.D. program: a 26-year experience. Acad Med. 1992;67:792–7.
19. Lynch KB, Woode MK. The relationship of minority students' MCAT scores and grade point averages to their acceptance into medical school. Acad Med. 1990;65:480–2.
20. Campos-Outcalt D, Rutala PJ, Witzke DB, Fulginiti JV. Performances of underrepresented minority students at the university of Arizona College of Medicine. Acad Med. 1994;69:577–82.
21. Lee MC. "Programming" minorities for medicine. JAMA. 1992;267:2391–4.
22. Glaser K, Hojat M, Callahan C. Evaluation of an enrichment programme for entering medical students predicted to be in need of academic preparation. Education for Health. 1996;9:221–8.
23. Watts VG, Harris CT, Pearson W. Course selections and career plans of black participants in a summer intervention program for minority students. Acad Med. 1989;64:166–7.
24. Julian TW, McKenry PC, Mckelvery MW. Cultural variations in parenting: perceptions of Caucasian, African-American, Hispanic, and Asian-American parents. Family Relations. 1994;43:30–7.

CLOSE BUT NO BANANAS: PREDICTING PERFORMANCE

Moderator: Mark Albanese, PhD

Does Institutional Selectivity Aid in the Prediction of Medical School Performance?

AMY V. BLUE, GREGORY E. GILBERT, CAROL L ELAM, and WILLIAM T. BASCO JR.

Various factors are considered in the decision to offer an admission interview to a medical school applicant, including Medical College Admission Test (MCAT) scores, undergraduate grade-point average (GPA), and the selectivity of the degree-granting undergraduate institution. Admission officers view MCAT scores, undergraduate GPA, and institutional selectivity as having high or moderate importance.[1] Research has indicated that these factors, most notably the MCAT scores and the undergraduate GPA, are reliable in helping predict medical school performance.[1–6] The strongest association has been shown between MCAT scores and performance on the United States Medical Licensing Examination, Step 1.[2]

Institutional selectivity data are used to help control for differences in grading stringency across undergraduate institutions.[1] Previous reports have examined the role of institutional selectivity, or a specific undergraduate institution, as a predictor of performance in the first two years of medical school.[1–3,6] With the exception of the study of Zelesnik et al.,[6] which examined ten specific undergraduate institutions, these reports have used the Higher Education Research Institute (HERI) Index,[7] also called the "Astin Index,"[2] as a measure of institutional selectivity. Other measures of institutional selectivity or categorization that schools of medicine may employ include the Barron's Profiles of American Colleges Admissions Selector Rating[8] and the Carnegie Classification from the Carnegie Foundation for the Advancement of Teaching.[9] (These measures are explained in the next section.)

Institutional validity studies of admission decision-making data help to determine which characteristics should be accorded highest importance in applicant selection. Given the reliance upon institutional selectivity as an important admission characteristic and the different types of selectivity classifications available for medical schools to use, the purpose of this study was to examine how well three measures of institutional selectivity could predict medical students' performances, specifically their performances on the USMLE Step 1 and Step 2 and their final medical school GPAs.

Method

Admission and medical school performance data were obtained for the 1992–1995 matriculants at the study institution, the Medical University of South Carolina (MUSC). Admission data for each student consisted of his or her MCAT scores, undergraduate GPA, undergraduate institution, three institutional selectivity categorization indices (the 1983 HERI index, Barron's Admissions Selector Rating, and the Carnegie Classification), age, gender, and underrepresented minority (URM) status. The 1983 HERI index consists of the mean total SAT score for all students admitted in 1983 to U.S. undergraduate institutions. The Barron's Profile of American Colleges Admissions Selector Rating indicates the degree of competitiveness of admission to a college.[8] The Carnegie Classification includes most colleges and universities in the United States that are degree-granting and accredited by an agency recognized by the U.S. Secretary of Education.[9] The Carnegie Classification is not meant to be a measure of selectivity. It is a classification of institutions into 19 categories based upon the ranges and types of degree-granting programs at the institutions (doctoral through associate of arts) and the amount of federal support annually received at each institution.

Medical school performance data consisted of USMLE Step 1 and 2 scores and final GPA. Students admitted under the institution's existing Early Assurance Program (EAP) were excluded from analysis because an MCAT score was not required for their admission. (The EAP offered admission to exceptional applicants during their undergraduate education based on the applicants' SAT scores, undergraduate GPAs, medical school admission interview ratings, and the understanding that the applicants would not apply to another medical school. This program stopped selecting applicants for admission to MUSC in 1996).

To avoid having insufficient subgroup size, we dichotomized the Barron's Admissions Selector Ratings and the Carnegie Classification categories based upon logical breakpoints in the categories. Calculated frequency distributions indicated that these breakpoints separated into approximately equal numbers of matriculants in each selectivity index or categorization grouping, thus confirming the breakpoints. The Barron's Admissions Selector Ratings were dichotomized into "most/highly competitive" (includes Barron's categories "most competitive," "highly competitive+" and "highly competitive") versus "not highly competitive" ("very competitive+," "very competitive," "competitive," "less competitive," and "not competitive"). The Carnegie Classification categories were dichotomized into either "research-doctoral" (includes Carnegie Classification Research I and II and Doctoral I and II institutions) and "not research-doctoral."

Descriptive statisrics were performed for all variables. Stepwise linear regression (adjusted r-square method) was used to assess which control variables (undergraduate GPA, gender, URM status, age) contributed significantly to predicting USMLE Step 1 and Step 2 scores and final GPA. Age was the only control variable that did not contribute significantly to predicting any of the dependent variables. Multiple linear regression was then performed with each of the institutional selectivity or categorization indices, controlling for URM status and gender. The powers of the multiple regression equations ranged from 88.2% to 96.0% for an alpha of 0.05 and with estimating of small effect sizes.

Results

For the 1992–1995 academic years, 545 applicants matriculated at MUSC. Of these, 112 were admitted under MUSC's Early Assurance Program (therefore missing MCAT scores) and were thus excluded from the study. Institutional selectivity index or categorization data were incomplete for an additional 28 matriculants, leaving complete data for 405 matriculants (73.3%).

Two hundred and sixty of the matriculants studied (64%) were men; 70 of the matriculants (17%) were from URM groups. The mean age was 24.0 years (SD = 4.0). The average total of MCAT subscores was 27 (SD = 4.2) and the average undergraduate GPA was 3.4 (SD = 0.40). Based upon the dichotomized Barron's Admissions Selector Rating, 235 of the matriculants (58%) had gone to undergraduate institutions that were classified as "not highly competitive." Using the dichotomized Carnegie classification, 233 of the matriculants (57%) had gone to research or doctoral undergraduate institutions. The mean USMLE Step 1 score was 205 (SD = 21), and the mean USMLE Step 2 score was 202 (SD = 21). The mean final medical school GPA was 3.3 (SD = 0.38).

Table 1 presents adjusted r-squared values for eight multiple regression models computed for the three dependent variables. All models predicted statistically significant variations in the dependent variables. Uniformly, the worst-fitting model was that which consisted of only the three control variables GPA, gender, and URM status. The amounts of explained variation ranged from 17% to 32%. Addition to the model of any institutional selectivity index or categorization slightly improved prediction (as measured by proportion of variation explained) above the prediction provided with GPA and demographic characteristics alone. When the MCAT score was added to the model involving the control variables and the GPA, it improved predictive ability of the equation by 6–13%. The addition of the institutional selectivity indices or categorization after the MCAT score was in the model added nothing to the predictive ability. Control variables plus MCAT score accounted for 38% of the variation in USMLE Step 1 scores, 38% of the variation in final GPA, and 28% of the variation in USMLE Step 2 scores.

Table 1. Percentages of Variation Accounted for in Predicting USMLE Step 1 and Step 2 Scores and Final Grade-Point Average with Three Institutional Selectivity Measures for 1992–1995 Medical University of South Carolina Matriculants

Model*	Percentage of Variation		
	USMLE 1 Score	USMLE 2 Score	Final GPA
GPA + gender + URM	25.18	17.22	32.18
GPA + gender + URM + Barron's selectivity index	25.96	19.45	33.77
GPA + gender + URM + Carnegie classification	26.40	17.83	32.85
GPA + gender + URM + Higher Education Research Institute selectivity index	26.27	18.14	33.04
GPA + gender + URM + MCAT	38.23	27.07	37.63
GPA + gender + URM + MCAT + Barron's selectivity index	38.25	27.80	38.51
GPA + gender + URM + MCAT + Carnegie classification	38.52	27.15	37.88
GPA + gender + URM + MCAT + Higher Education Research Institute selectivity index	38.24	27.08	37.81

* GPA = undergraduate grade-point average; gender = man or woman; URM = underrepresented minority; Barron's selectivity index = Barron's Profile of American Colleges Admission Selector ratings dichotomized into "most/highly competitive" versus "not highly competitive"; Carnegie classification = Carnegie Foundation for the Advancement of Teaching Classification dichotomized into "research-doctoral" versus "not research-doctoral"; Higher Education Research Institute selectivity index = mean total SAT score for all students admitted in a given year at a particular institution.

Discussion

During the medical school admission process, the selectivity of the degree-granting undergraduate institution is used to help control for grading differences across undergraduate institutions. Our results show that none of the three institutional selectivity indices or categorizations (i.e., HERI, Barron's, or Carnegie) and any GPA adjustment that would follow will improve correlation with performances on USMLE Step 1 and Step 2 and final GPA if MCAT scores and unadjusted GPA are used in conjunction. While the Barron's and HERI indices are meant to be measures of institutional selectivity, the Carnegie classification is a description of the degree spectrum offered. Even evaluating schools by the type of degree offered produced no added benefit to prediction.

Previous studies have shown that selectivity measures aid prediction of the USMLE Step 1 score and the GPAs in medical school years one and two if used in a model without the MCAT scores. However, those studies used only one measure of institutional selectivity, the HERI,[2,3] or a sampling of undergraduate institutions.[6] Our study evaluated three different methods of classifying the selectivity or type of undergraduate institution, and none improved prediction in models that included the MCAT score. Furthermore, our study examined performance on USMLE Step 2 and final medical school GPA, performance indicators beyond the first two years of medical school.

Our findings suggest that using institutional selectivity indices or categorizations as an admission characteristic may not be necessary. In addition, use of institutional selectivity indices or categorizations may discriminate against applicants with other desirable characteristics who have been granted degrees from less selective undergraduate institutions. For example, use of the average SAT score might unfairly discriminate against applicants who graduated from large, state-sponsored universities. The lack of correlation with the Carnegie classification also indicates that the size or academic comprehensiveness of the degree-granting institution has little bearing on individual performance in medical school. Our results should reassure admission officers that the performances of students who attend smaller undergraduate institutions or community colleges are predictable when using their MCAT scores and undergraduate GPAs.

One limitation of this study is that it relied upon data from only one, state-supported, medical school. However, matriculants at the school come from diverse undergraduate institutions, 116 for the individuals in this study. Additional research should examine this issue at other medical schools, both state-supported and private and in various regions of the United States. Another limitation is that because multiple linear regression was used, correlations with USMLE scores and final GPAs cannot be adjusted for restriction in range. Thus, the adjusted r-square values presented in Table 1 are, in all likelihood, underestimates of the relationships between the models and the dependent variables for the applicant pool. In addition, the dichotomization of the Barron's Admissions Selector Ratings and the Carnegie Classification categories may also have had some impact on our results. However, any contributed bias would likely have strengthened the ability of institutional selectivity to influence the performances of students. Another limitation is that the HERI index, although the most recent currently available, is quite dated (1983); hence, the HERI index may not be representative of today's undergraduate institutions. Finally, this study focused on primarily cognitive measures of academic achievement in medical school. The predictive value of institutional selectivity indices or categorization on performances in clinical settings also should be explored.

In summary, our results indicate that the characteristics of the degree-granting undergraduate institution, as measured by three different types of institutional selectivity or categorization, do not add to the ability to predict performances on USMLE Steps 1 and 2 and overall medical school GPA if the MCAT score and unadjusted undergraduate GPA are available. The results also further support the predictive validity of the scores on the MCAT examination for medical school performance.

Correspondence: Amy V. Blue, PhD, College of Medicine, Dean's Office, Medical University of South Carolina, 96 Jonathan Lucas Street, Suite 601, Charleston, SC 29425; e-mail: (blueav@musc.edu).

References

1. Mitchell K, Haynes R, Koenig J. Assessing the validity of the updated Medical College Admission Test. Acad Med. 1994;69:394–401.
2. Wiley A, Koenig J. The validity of the Medical College Admission Test for pre-

43

dicting performance in the first two years of medical school. Acad Med. 1996;71(10 suppl):S83–S85.
3. Swanson DB, Case SM, Koenig J, Killian CD. Preliminary study of the accuracies of the old and new Medical College Admission Tests for predicting performance on USMLE Step 1. Acad Med. 1996;71(1 suppl):S25–S27.
4. Mitchell KJ. Traditional predictors of performance in medical school. Acad Med. 1990;65:149–58.
5. Huff KL, Fang D. When are students most at risk of encountering academic difficulty? A study of the 1992 matriculants to U.S. medical schools. Acad Med. 1999; 74:454–60.
6. Zelesnik C, Hojat M, Veloski, JJ. Predictive validity of the MCAT as a function of undergraduate institution. J Med Educ. 1987;62:163–9.
7. Higher Education Research Institute. UCLA Graduate School of Education and Information Studies [unpublished data].
8. Barron's Profiles of American Colleges, 23rd ed. Hauppauge, NY: Barron's Educational Series, July 1998.
9. Boyer E. A Classification of Institutions of Higher Education. Pittsburgh, PA: The Carnegie Foundation for the Advancement of Teaching, 1994.

The Presence of Hospitalists in Medical Education

JUDY A. SHEA, YASMINE S. WASFI, KIMBERLY J. KOVATH, DAVID A. ASCH, and LISA M. BELLINI

Over the past few years, the care of medical inpatients increasingly has been turned over to an emerging group of professionals called *hospitalists*. Typically, hospitalists are internists who devote the majority of their clinical effort to caring for inpatients.[1-4] Outpatient responsibilities, if they exist, are minimal. The defining characteristic of hospitalists is the "hand-off" cycle: a primary care provider admits the patient to the designated hospitalist, who provides inpatient care and then sends the patient back to the primary care provider upon discharge.

Many early arguments for hospitalists centered on the positive impact that this model of care would have on resource utilization and patient outcomes. Indeed, early data suggest that care by hospitalists is associated with reductions in length of stay, lower readmission rates, and improved resource utilization,[5-8] and there seems to be little negative impact on patients' satisfaction.[9] Among the issues that have not been fully addressed is the role that hospitalists play in medical education. Potential issues have been discussed, such as a diminished sense of autonomy among residents,[10,11] perhaps counterbalanced by increased satisfaction and better supervision of patients.[4,9] For other issues, such as the presence of hospitalists in academic medical centers and their teaching responsibilities, few data have been presented. Historically, the cornerstone of both undergraduate and graduate medical education has been inpatient-based. Though ambulatory care training has been emphasized in recent years, the inpatient wards remain the major site of clinical teaching. If beds and/or wards are being turned over to hospitalists, it is important to determine the impact this may have on educational programs.

The purpose of this study was to address such educational issues. Separate questionnaires were mailed to the chairs and program directors of all internal medicine training programs in the United States to learn (1) how many programs have hospitalists on staff, and to gain information about related census issues (e.g., number hired, plans for future hires); (2) the role of hospitalists in teaching activities, and (3) attitudes regarding the roles hospitalists play in general and their role in teaching, specifically.

Method

The questionnaire was sent to all chairs and program directors of accredited internal medicine training programs who were identified in the spring of 1999 using the 1998–1999 AMA *Graduate Medical Education Directory*. This process resulted in a roster of 106 chairs, 382 program directors, and 22 individuals who filled both roles. Three separate questionnaires were developed. Content was defined by the study team, taking ideas from current literature as well as discussions that had taken place locally in the course of developing a hospitalist service in 1998. Draft instruments were revised numerous times to improve clarity and breadth, after piloting them with faculty.

The questionnaire for chairs was brief (eight questions) and focused on asking whether hospitalists were employed at the sites and if so, defining how long they had been there and their training backgrounds and responsibilities. A more extensive questionnaire was developed for program directors. In addition to general program descriptions, directors were asked whether hospitalists were employed and provided 12 attitude statements about hospitalists to be answered on a five-point Likert scale from "strongly disagree" to "strongly agree." For programs that had hospitalists on staff, a set of questions focused on teaching responsibilities, participation in other educational activities, and 13 more attitude statements about hospitalists' roles and their impact upon teaching. The questionnaire for the few individuals who were both chairs and program directors was a collection of the unique items from the other two versions. The first mailings were sent in April 1999. A second mailing with a new copy of the instrument was sent in June 1999. Because the response rate for program directors was low, for the third mailing, items asking about activities at each training site were omitted to reduce the respondents' burden, thus shortening the questionnaire from four to two pages. The third mailing was sent in August 1999. The final response rates were 78.3% (n = 83) for chairs and 57.6% (n = 220) for program directors. The eight responses from the 22 chairs–program directors were added to both data files, for analytic sample sizes of 91 and 228.

Analyses of the responses focus on description. We used standard univariate statistics (frequencies and percentages) to characterize the sample. To test for differences between programs that did and did not respond, between responses to the long and short survey forms, and between programs that did and did not employ hospitalists, we used chi-square, t-tests, and the Wilcoxon two-sample test.

Results

Respondents and Non-respondents. Program characteristics available from the 1998–1999 AMA *Graduate Medical Education Directory* allowed limited comparison of non-respondents with respondents. Overall, the program sizes were the same for respondents (mean = 52.9, SD = 33.2) and non-respondents (mean = 55.8, SD = 36.9, p = .40). The respondents and the non-respondents did not come from different regions of the country (p = .086).

There were few significant differences between the responses of the 130 program directors who responded to the long form of the program directors' questionnaire and the 90 who responded to the short one. For example, there was no difference in numbers of inpatient training sites (p = .63) or numbers of categorical residents at the PGY1 level (p = .35). There was no difference in the percentages who had hospitalists (p = .80), were planning to hire hospitalists (p = .59), or had rejected the idea of having hospitalists (p = .47). Those responding to the short form had more favorable attitudes with respect to one of the 13 attitude items.

Chairs. Overall, 50 of the chairs (55.6%) reported that hospitalists were employed at one or more of their training sites. The numbers of hospitalists per institution ranged from one to 15, with a median of four. (The total number of hospitalists employed by the 44 programs that reported having them was 206.5.) Twenty-nine (64.4%) planned to hire more hospitalists. The tenure of hospitalist programs was a median of two years, with a range of 0.5 to 7.5 years. Nearly three fourths of the hospitalists (71.9%) had completed residencies in internal medicine, 4.4% had completed general internal medicine fellowships, and 11.4% had completed subspecialty fellowships.

The reported duties of the hospitalists were quite variable. The numbers of months of inpatient responsibilities ranged from one to 12, with a median of eight. The percentages of the responding department chairs reporting other responsibilities for hospitalists

45

TABLE 1. Attitudes of IM Program Directors Regarding Hospitalists, and Comparison of Attitudes of Those Whose Programs Do and Do Not Employ Hospitalists, 1999

Questionnaire Statement	All Program Directors* (N = 217) % Disagree	% Neutral	% Agree	Comparison of Two Groups of Program Directors† Those with Hospitalists (n = 109)	Those without Hospitalists (n = 108)	p
Hospitalists are more familiar with practical aspects of inpatient care than other physicians attending on inpatient services.	23.9	17.7	58.4	3.55	3.25	.04
Hospitalists need additional training beyond standard internal medicine residency.	53.1	27.1	19.8	2.57	2.77	.39
Hospitalists provide better inpatient care than other general internists attending on inpatient services.	33.3	33.3	33.3	3.03	2.94	.61
Hospitalists provide better inpatient care than subspecialists attending on inpatient services.	33.0	27.3	39.7	3.13	2.90	.13
Patients of hospitalists are satisfied with the inpatient care they receive.	3.5	42.6	54.0	3.81	3.24	.0001
Patients of hospitalists are satisfied with the outpatient care they receive.	4.5	61.6	33.8	3.35	3.23	.18
The use of hospitalists disrupts continuity of patient care.	25.8	20.1	54.1	3.21	3.51	.17
The use of hospitalists improves patient care.	11.1	47.1	41.8	3.49	3.18	.005
The use of hospitalists is good for hospitals financially.	2.4	40.1	57.5	3.83	3.52	.008
The use of hospitalists improves the training of residents.	17.7	45.5	36.8	3.45	2.99	.0003
The use of hospitalists improves the training of medical students.	18.8	46.6	34.6	3.38	3.01	.004
I expect that use of hospitalists will increase over the next few years.	3.8	11.0	85.2	4.16	3.86	.0025

*Responses were made on a five-point scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. In reporting the overall responses, the 1 and 2 categories were collapsed, as were the 4 and 5 categories.

†Group comparisons were made between program directors from programs with and without hospitalists. Scores on the five-point scale were compared with the Wilcoxon two-sample test.

were: 55.3% reporting hospitalists with outpatient practices (with a median of 10% full-time equivalent); 77.8%, medicine consultation; 46.8%, clincial pathways/disease management development or implementation; 31.9%, quality assurance; 27.7%, medical directorships; and 17.4%, insurance company or managed care liaisons. Of note, 53.2% required academic productivity for promotion.

Of the programs that did not have hospitalists, 37.1% planned to hire them in the future and 16.1% had considered but rejected the idea.

Program Directors. Overall, 50.5% of the responding programs employed hospitalists. As shown in Table 1, many program directors' attitudes about hospitalists were positive. For example, the majority agreed that hospitalists are more familiar with practical aspects of inpatient care, that they are good for the hospital financially, and that patients of hospitals are satisfied with their inpatient care. Most disagreed that they needed more training beyond that gained in an internal medicine residency. On the other hand, most also thought that use of hospitalists disrupted the continuity of patient care, and only one third agreed that hospitalists provide better inpatient care than other general iternists.

The last three columns of Table 1 shows the means of the Likert-scale responses of the program directors with and without hospitalists. Differences were significant, and in the anticipated direction, for seven of the 12 attitude statements.

In addition, respondents of the 109 programs with hospitalists were asked whether the hospitalists participated in a number of different activities related to education. Nearly all participated in the teaching of medical students (80.2%) and residents (84.5%). Other educational activities in which they participated included attending physicians' rounds (74.7%); residents' reports (58.6%); management conferences (53.5%); curriculum development (55.6%), and journal club (48.5%). Specific topics taught by the hospitalists included cost-effective care (57.1%); resource utilization (57.1%); health economics (42.9%); clincial pathways/disease management (38.8%); and insurance principles (26.5%). In nearly all programs (78.1%), students and housestaff evaluated the hospitalists.

Table 2 lists additional attitudes of the program directors who employed hospitalists, especially their perceptions of hospitalists' role in and impact on teaching activities. Over 70% agreed that that hospitalists are viewed as good educators and are respected. The majority thought thar hospitalists have led to improved housestaff supervision and are more accessible to housestaff than other teaching faculty. They were less certain that hospitalists had an impact on housestaff's considerations of lengths of stay and costs of tests and procedures, or that the housestaff had learned to order fewer tests and procedures.

Discussion

The results of the surveys reported above show that hospitalists have a presence in both undergraduate and graduate medical education: at least half of the responding training programs employed hospitalists, who in most cases played roles in teaching students and/or residents. Attitudes expressed by the total sample of program directors were generally positive, naturally more so for those representing programs with hospitalists. In particular, program directors from programs with hospitalists were especially complementary about the hospitalists' familiarity with practical aspects of care, their positive financial impact on the hospitals, their positive impact on patients' satisfaction, and the improvements in residency training. On the other hand, most programs had only a few hospitalists, they had had them for only one or two years, the numbers of months in inpatient responsibility ranged from one to 12, and their involvement in a variety of specific teaching activities was varied. Given this variation, it might not be feasible to characterize "the" teaching role of hospitalists.

This study has some limitations. The response rate for program directors was relatively low, although we found no evidence of bias. Second, we are able to create a composite picture of what hospitalists do, but we did not collect parallel data regarding non-hospitalist attending physicians. Thus, we are missing a piece of the total picture. Third, we did not ask for detailed data regarding the teaching activities of the hospitalists, e.g., what does participation

Table 2. Attitudes Held by 109 Directors of IM Programs That Employ Hospitalists

Questionnaire Statement	% Disagree*	% Neutral	% Agree*
The hospitalists at any institution are viewed as good educators.	3.2	18.3	78.5
Housestaff supervision has improved with the addition of hospitalists.	17.8	30.0	52.2
Housestaff have adequate exposure to physician–scientist faculty.	21.1	25.0	57.6
Hospitalists are more accessible to housestaff than other teaching faculty.	21.1	16.7	62.2
Housestaff are more comfortable managing inpatients with hospitalists than with other general medical attendings.	27.8	31.1	41.1
Housestaff are more comfortable managing inpatients with hospitalists than with subspecialist attendings.	35.6	36.7	27.8
Hospitalists are respected at my institution.	6.4	21.3	72.3
Housestaff who have worked with hospitalists consider length of stay in their management plans more than they did previously.	20.0	42.2	37.8
Housestaff who have worked with hospitalists consider cost of tests and procedures when determining their management plans more than they did previously.	18.7	44.0	37.4
Housestaff-perceived autonomy has been reduced by the use of hospitalists.	47.3	29.7	23.1
The use of hospitalists has reduced the inpatient teaching responsibilities of the other faculty physicians.	31.9	12.1	56.0
The use of hospitalists as teaching attendings has resulted in less interaction between housestaff and primary care providers.	44.9	25.8	29.2
Housestaff who have worked with hospitalists have learned to order fewer tests and procedures than they did previously.	22.5	52.8	24.7

*Responses were made on a five-point scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. The 1 and 2 categories were collapsed, as were the 4 and 5 categories.

in curriculum development mean? Nevertheless, to our knowledge this is the first study to detail the presence of hospitalists and provide an overview of their teaching activities in teaching institutions.

Overall, even though their numbers were small, in at least half of the U.S. internal medicine training programs that responded, hospitalists were present and played roles in teaching. Given the amount of time they spend in inpatient services, they have widespread exposure to learners on all levels. This visibility makes hospitalists as a group an ideal target for faculty development focused on teaching methods and feedback skills. Although the respondents generally viewed hospitalists as excellent teachers who had led to improved training for residents, hospitalists as teaching faculty should be evaluated compared with faculty involved in teaching on traditional services.

Much of the justification for hospitalists draws on arguments that they should be able to save money by reducing test ordering and lengths of stay. If they are really succeeding in these areas, as current data suggest they might be,[5–8] it is logical to assume they could affect residents' behaviors via modeling and/or direct teaching of optimal management strategies. The fact that so few program directors believe hospitalists will have an effect on residents' future behaviors in the areas of ordering and effective management is somewhat surprising and deserves more study. Examining residents' behaviors and attitudes in terms of the relative amounts of exposure they have to services led by hospitalists would yield useful insights. An additional contribution would be an understanding of how, as a group, hospitalists' teaching activities and outcomes differ from those of their peers, and whether there is observable variation in the reasons hospitalists' services were developed (e.g., cost efficiency, excellence in inpatient teaching, as a "safety valve" for overburdened teaching services). Future studies will be most helpful if they are aimed at defining the unique contributions attributable to hospitalists.

Correspondence: Judy A. Shea, PhD, University of Pennsylvania, 1232 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021; e-mail: ⟨sheaja@mail.med.upenn.edu⟩.

References

1. Wachter RM. An introduction to the hospitalist model. Ann Intern Med. 1999;130(4:part 2):338–42.
2. Wachter RM, Goldman L. The emerging role of hospitalists in the American health care system. N Engl J Med. 1996;335:514–7.
3. Weissler JC. The hospitalist movement: caution lights flashing at the crossroads. AJM. 1999;107:409–13.
4. Whitcomb WF, Nelson JR. The role of hospitalists in medical education. Am J Med. 1999;107:305–9.
5. Craig E, Hartka L, Likosky WH, Caplan WM, Litsky P, Smithey J. Implementation of a hospitalist system in a large health maintenance organization: the Kaiser Permanente experience. Ann Intern Med 1999;130(4:part 2):355–9.
6. Freese RB. The Park Nicollet experience in establishing a hospitalist system. Ann Intern Med. 1999;130(4:part 2):350–4.
7. Diamond HS, Goldberg E, Janosky JE. The effect of full time faculty hospitalists on the efficiency of care at a community teaching hospital. Ann Intern Med. 1998;129:197–203.
8. Brown MD, Halpert A, McKean S, Sussman A, Dzau VJ. Assessing the value of hospitalists to academic health centers: Brigham and Women's Hospital and Harvard Medical School. Am J Med. 1999;106:134–7.
9. Wachter RM, Katz P, Showstack J, Bindman AB, Goldman L. Reorganization of the academic medical service: impact on cost, quality, patient satisfaction and education. JAMA. 1998;279:1560–5.
10. Goldman L. The impact of hospitalists on medical education and the academic health center. Ann Intern Med. 1999;130(4:part 2):364–7.
11. Schroeder SA, Schapiro R. The hospitalist: new boon for internal medicine or retreat from primary care? Ann Intern Med 1999;130(4:part 2):382–7.

47

Dual-degree MD–MBA Students: A Look at the Future of Medical Leadership

WINDSOR WESTBROOK SHERRILL

In an increasingly turbulent medical care system, business training is one way doctors and medical students are seeking to redefine their ability to lead and wield influence. Changes in the health care system have fostered the need for physician executives with business training who can serve as liaisons between administrative and clinical personnel. As the development of integrated delivery systems has combined clinical and administrative functions, the roles of physician executives have increased, as well as the demand for related training of physician leaders. Growth in the number of physician executives is expected to continue as such individuals demonstrate their ability to facilitate provider–physician relations and lend unique expertise and perspective in the health care delivery system.[1]

The transition from clinical roles to administrative functions can be challenging for physician executives.[2,3] Moving into administrative roles presents challenges different from those inherent in medical training and practice.[4] If the physician manager is to be considered an effective asset to an organization, the new role requires distinct shifts in thinking, philosophy, attitudes, and behavior.[5] Because traditional clinical training of physicians contrasts with management training and functions, few physicians are prepared for the requirements of management roles.[6]

Several studies of leadership and management have found that leaders' personality and behavioral characteristics are reliably predictive of group performance.[7,8] Leadership success is associated with interpersonal ability, group-oriented behaviors, empathy, boldness in times of uncertainty, internal locus of control, and confidence.[9,10] Leadership theory suggests that effective leaders are able to identify and actively respond to changes in a profession, organization, or situation.[8]

Although a growing number of practicing physicians have obtained business (MBA) degrees, relatively few educational initiatives have been focused on business and management training within the medical school program.[11] In response to this demand, a limited number of medical schools are offering dual-degree programs in medicine and business. Established through cooperative agreements between medical and business schools, these programs offer a variety of arrangements through which medical students can obtain business and clinical training concurrently.

Students enrolled in dual-degree programs make up an important group for exploratory research. If dual-degree medical students exhibit characteristics associated with successful leaders, this might indicate their ability to function as effective leaders in both clinical and management roles. Within the traditional medical school environment, it is possible that this group is reshaping individual beliefs about physician roles and the fit between clinical and administrative functions. Their career goals and the factors influencing these students to seek business training might provide an indicator of the leadership styles and roles of future physician executives.

Method

According to *Peterson's Guide to Graduate Programs in Business, Education, Health and Law*, there were eight medical schools that offered dual-degree MD–MBA programs in 1997. Of the eight schools, six had coordinated MD–MBA programs for which program directors were designated and students followed a defined path in course work. Students in these programs were selected for inclusion in the present study.

Of the six dual-degree programs, one MD–MBA program could be completed in four years by using summers for course work. The other programs that were examined required five or six years of study. Each of the dual-degree programs had some component of an administrative internship for the students.

Survey and interview measures were used to analyze students at the six medical schools offering MD–MBA programs ($n = 87$): Bowman Gray School of Medicine, Jefferson Medical College, the University of Chicago, the University of Pennsylvania, the University of Illinois at Urbana–Champaign, and Tufts Medical School. The 87 students who were enrolled in dual-degree programs were surveyed; a control group of traditional medical students was also surveyed ($n = 115$). Traditional medical students at each site were selected based on a set of characteristics matched with those of the dual-degree students. The data were also compared with the findings from a national survey of graduating medical students compiled by the Association of American Medical Colleges. Forty of the 87 dual-degree medical students surveyed were also interviewed. The interviews were analyzed using Ethnograph, and survey data were analyzed using SPSS.

To assess whether they might overcome the barriers between clinical and management roles, dual-degree medical students were compared with traditional medical students on dimensions that were selected for their potential to indicate leadership ability. Dual-degree students were also asked about their career goals and the factors that had influenced them to seek business training.

Results

The response rate for the survey of the 87 MD–MBA students was 85%; the response rate for the 115 medical students in the control group was 69.6%. A major finding of the study is that there are indeed significant differences between dual-degree and traditional medical students on a number of dimensions that relate to career plans, leadership, motivation to be leaders, and confidence.

One set of questions was intended to assess students' beliefs, concerns, and perceptions about the future of medical practice. The set of questions was designed to compare the attitudes of dual-degree and traditional medical students regarding the changes in health care and the evolution of the physician role. The students were asked to rank statements such as "job opportunities for physicians are increasingly limited" and "the health care financing system is too burdensome on physicians." Answers to these questions provided a composite index of students' perceptions, including attitudes concerning the role of physicians in society. Both survey responses and interview feedback support the hypothesis that dual-degree students are very conscious of the changing nature of the medical care system and the need to transform physician roles. Dual-degree students were less likely to feel negative about changes in job opportunities for physicians or about regulatory or financial constraints in medicine. The data also support the hypothesis that dual-degree students are influenced to obtain business degrees because of concern about the changing job market for physicians.

The members of both the dual-degree and control groups were asked what they expected to earn five and ten years after completing residencies. The MD–MBA group had an expected mean in-

come after five years of $167,986, while the MD students had a mean of $132,208. The means of the two groups were significantly different; $t\ (147) = 3.66$, $p < .0001$.

As an indicator of their career plans and aspirations, dual-degree students were asked to rank activities according to how they would feel about them as primary job responsibilities. Job responsibilities ranged from CEO of a for-profit hospital to medical director of an inner-city health clinic. Job responsibilities were provided as indicators of the types of positions these students might desire, particularly related to their tendencies toward more altruistic positions compared with activities that might be traditionally associated with the "business" of medicine. The job activities were organized into subgroups based on activity type and were developed to reflect items that might indicate students' altruistic versus economic philosophies. The first group included medical director of an HMO, CEO of a biotechnology company, medical director of an insurance company, and chief of staff of a for-profit hospital system. In contrast to the first group of activities, the second group included activities traditionally associated with the public-services arena. This group included medical director of an inner-city health clinic, chief of staff of a rural hospital, medical liaison for the World Health Organization, and deputy director of the state board of health.

The combined group ratings were compared using t-tests, and the subgroup scores were significantly different. The mean for the "business" subgroup was 1.83; the "public services" subgroup mean was 2.26. The dual-degree students considered the business group significantly more appealing, $t\ (105) = 3.02$, $p < .05$.

Both dual-degree and traditional medical students were asked to select their preferences from a list of career activities, including such things as full-time faculty appointment, private clinical practice, and administrative duties. Seventy-eight percent of the dual-degree students expressed an interest in a combination of clinical and administrative duties; 13.5% of the dual-degree students planned administrative jobs with no clinical practice. Several dual-degree students interviewed stated that they planned to forego residency training to initiate careers in the private sector.

Both traditional medical and dual-degree students were asked whether they were confident that they would have necessary clinical and administrative skills when they graduated from their respective educational programs. These results were compared with corresponding information from the national database of graduating medical students as well as from the control group of students. Dual-degree students expressed little doubt in their clinical or administrative abilities and were significantly more confident than were their medical student counterparts (clinical skills—$t\ (151) = 6.409$, $p < .0001$; administrative skills—$t\ (150) = 2.913$, $p < .01$).

Confidence in one's ability to influence others and the environment is associated with leadership.[12] Yet, misplaced confidence can lead to poor decision making for both clinicians and managers. It is interesting that the dual-degree students were more confident than were the traditional medical students with respect to both clinical and administrative skills. Although a positive self-concept may be beneficial, the students' confidence has implications for the future of medical care. The potential overconfidence of the MD–MBA students needs to be understood and managed to avoid potential disastrous effects; confidence is a positive attribute for leaders and managers, but overconfidence can be a barrier to effective decision making. Individuals who are overconfident might fail to seek consensus among groups and lack the discipline to seek out information in solving clinical and management problems.

Students' influences and motives for choosing the dual-degree programs, as well as their career plans, provide an indication of the roles they will play in the delivery system. It was hypothesized that dual-degree students are motivated to seek business degrees because of a desire to be leaders in the health care delivery system. Their career goals and plans illustrate such motivation. In response to survey questions related to their reasons for seeking business education, the students rated most highly factors such as career opportunities, opportunity for innovation, opportunity to be a leader in medicine, and opportunity to make a difference in medicine.

Discussion and Implications

A new model of physician executives is emerging from dual-degree programs. Young physicians are making decisions not only at the beginning of their medical careers, but in most cases, for these students, at the beginning of their medical education. It is possible that dual-degree programs will produce individuals equipped to take leadership roles in managerial assignments early in their careers, perhaps even in residency programs.

This study underscores an important policy question for the health care system and medical education. The challenges facing the health care system are both economic and equity-related. Escalating costs are combined with serious problems of underserved populations. Some of the most significant management challenges in the delivery system relate to the challenge of how to provide equalized distribution of health care services as well as how to improve access to basic health care services. Physicians with business training are needed in all areas, not only in the areas of high technology and high costs. The study findings suggest that dual-degree programs are attracting students primarily with business interests. Eleven of the 40 students interviewed had full-time and significant work experience in areas such as investment banking and health care consulting prior to matriculation in medical school. Students interested in working with public health needs and underserved populations are not well represented in the dual-degree programs.

Are the programs too narrowly focused on dealing with the business of health care delivery? As early adopters of an innovative medical education initiative, dual-degree students provide a unique perspective on the direction of medical leadership and alternatives to traditional medical careers. A key finding of the study suggests that this cohort wants to direct hospital and insurance companies more than they want to work in the public sector. This indicates that the motivation for those students to seek dual-degree programs, as well as motivating factors behind program development, were related to business and high-technology settings. The students' job-activity preferences and income expectations provide support for this conclusion. Traineeship experiences and mentors provided by the dual-degree programs may need to be modified to address these trends.

Physician executives are likely to have a pivotal role in the uncertain future of health care.[13] The management of health care resources requires a combination of skills that balance the principles of economics, finance, and accounting with patient and population health needs. Dual-degree medical education programs can help develop physician leaders who can blend clinical and management skills into an effective vision for the future of health care delivery.

The authors of *In Search of Physician Leadership* observe that physicians are entering management in increasing numbers and at increasing levels of responsibility, a trend they assert portends well for the medical profession and the health care system.[14] Medical education programs combining business and clinical education are training students who can contribute to this positive trend. As one student stated, "[Dual degrees] can bring values of medicine into the business world. It used to be totally different, but now things are beginning to merge. We can do the best for both fields."

This is an exploratory study of an innovation in medical education. The early stages of this field offer the opportunity to step back and consider the professional identity desired among dual-degree medical students. Dual-degree programs are producing a prototype of physician executive whose training is remarkably different from that of traditional physicians. The data suggest that there is an interesting range of expectations among dual-degree medical students and the careers that they anticipate. The interests and career

49

preferences of the students reveal several trends of concern, but also suggest that these programs can make an important contribution to the health care system.

Correspondence: Windsor Westbrook Sherrill, MHA, PhD, Assistant Professor, Department of Public Health Sciences, 525 Edwards Hall, Clemson University, Clemson, SC 29632; e-mail: (*Wsherri@clemson.edu*).

References

1. Smallwood KG, Wilson CN. Physician-executives past, present and future. South Med J. 1992;85:840-4.
2. Peters RM. When Physicians Fail as Managers: An Exploratory Analysis of Career Change Problems. Tampa, FL: American College of Physician Executives, 1994.
3. Curry Wesley (ed). Roads to Medical Management: Physician Executives' Career Decisions. Tampa, FL: American Academy of Medical Directors, 1988.
4. Hagland MM. Physician executives bring clinical insight to non-clinical challenges. Hospitals. September 20, 1991.
5. Kurtz M. The Dual Role Dilemma in New Leadership in Health Care Management. Tampa, FL: American College of Physician Executives, 1992.
6. Ott JE. Administrative medicine. JAMA. 268;3:332-3.
7. Chemers MM. An integrative theory of leadership. In: Leadership Theory and Research. San Diego, CA: Academic Press, 1993.
8. Avolio BJ, Bass BM. Transformational leadership: a response to critiques. In: Leadership Theory and Research. San Diego, CA: Academic Press, 1993.
9. Bass BM. Leadership Performance Beyond Expectations. New York: Free Press, 1985.
10. House RJ, Shamir B. Toward the integration of transformational, charismatic and visionary theories. In: Leadership Theory and Research. San Diego, CA: Academic Press, 1993.
11. Wholey MH, Chapman JE. Business and managerial education in the medical school curriculum. South Med J. 1990;83:204-5.
12. Sashkin M, Burke WW. Understanding and assessing organizational leadership. In: Measures of Leadership. West Orange, NJ: Leadership Library of America, 1990.
13. Enthoven AC, Vorhaus CB. A pivotal role for physician executives. Physician Executive. 1990 July; 16:6-7.
14. LeTourneau B, Curry W. In Search of Physician Leadership. Chicago, IL: Health Administration Press, 1998.

50

A Preliminary Analysis of Different Approaches to Preparing for the USMLE Step 1

RAJ A. THADANI, D. B. SWANSON, and ROBERT M. GALBRAITH

Prior to taking the United States Medical Licensing Examination (USMLE) Step 1, medical students commonly spend large amounts of time studying on their own or in groups. They also may participate in test-preparation activities offered by their schools. In addition, there is anecdotal evidence indicating that medical students increasingly purchase "board prep" publications and sign up for commercial coaching courses that may last several weeks and cost thousands of dollars.

The effectiveness of alternate approaches to preparing for Step 1 is unknown, though research on other high-stakes exams suggests that exam performance may be improved by coaching courses.[1-4] In this article, we report the results of a recent survey of the strategies used by medical students to prepare for Step 1, and present our preliminary analyses of the relationships between preparation strategies and test scores.

Method

Participants in this study were chosen by taking a random sample of first-time Step 1 takers from U.S. and Canadian allopathic medical schools. In a survey following the June 1998 paper-and-pencil administration of Step 1, participants were asked about their study habits in relation to: the number of hours spent studying for Step 1 each week; the types of materials they had used when studying; the number of weeks of full-time study; and if applicable, a series of questions about coaching course(s). Survey responses from each participant were matched to his or her USMLE Step 1 scores, Medical College Admission Test (MCAT) scores, and undergraduate science grade-point average (GPA).

The survey questionnaires was mailed to a random sample of 3,958 first-time takers of Step 1. A total of 1,650 responded, but because MCAT scores and GPAs were not available for all examinees; only 1,217 were included in the analysis reported here. Initially, information about all the test-preparation materials and courses (methods) was examined descriptively. Then, to evaluate the usefulness of the various preparation methods, ordinary least-squares (OLS) regression analysis was employed. The variables included in the equation fell into four sets.

- *Pre-matriculation Characteristics.* MCAT total score (sum of biological sciences, physical sciences, and verbal reasoning scores) and adjusted science GPA.[5] Scores were calculated using the following formula: *adjusted science* GPA = *undergraduate science* GPA × *selectivity Index* ÷ *1,000*, where the selectivity index is equal to the mean Scholastic Aptitude Test score for students at the undergraduate school attended. This adjustment controls, to a degree, for the variation in grading stringency across undergraduate institutions.
- *Study time.* Weeks of full-time study and hours studied per week.
- *Preparation methods.* Use of USMLE materials, lecture notes, course syllabi, note-taking service, textbooks, commercial study guides, school materials, group study.
- *Coaching course.* A 0/1 dummy code that reflects participation in a coaching course plus an interaction term equal to the product of the dummy code and the MCAT total score.

Results

Descriptive Information about Preparation Methods. Analysis of the responses to the survey showed that 98% of the respondents had used commercial guides. In contrast, only 70% had used the official *USMLE General Instructions, Content Description and Sample Items Booklet*, which was provided to all examinees with application materials. Other methods had been used less often: lecture notes (39%), note-taking services (6%), textbooks (44%), course syllabi (21%), preparation materials provided by the school (25%), and group study (25%).

The survey responses also indicated that 23% of the respondents had enrolled in a commercial coaching course. When asked about the emphasis of the coaching course, 57% of these respondents reported that it had focused on learning Step 1 content, but also included instruction on test-taking strategies. Twenty-eight percent indicated that the emphasis of their courses had been entirely on learning Step 1 content. Less than 5% reported that their courses had spent a majority of the time on test-taking strategies. Examinees were asked to subjectively rate the value of their courses on a scale of 1 to 5, with 1 being not helpful and 5 being very helpful. The examinees' average rating was 3.2.

Analysis of mean scores attained by users versus non-users of certain study methods showed significantly better performances among examinees who had used the USMLE general instructions booklet, textbooks, course syllabi, and study materials provided by the medical school. Examinees who had enrolled in coaching courses received on average lower scores than those who did not enroll (Table 1). However, it should be noted that the examinees who had enrolled in coaching courses had significantly lower MCAT scores (28.8 versus 30.2) and adjusted science GPAs (3.70 versus 3.88). Differences in mean scores between users and non-users of other test preparation methods were smaller and statistically insignificant.

Descriptive Information about Study Time. In order to examine the effects of preparation time, the examinees were asked how many weeks they had spent studying for the exam full time and how many hours per week they studied during that time. Since these examinees were first-time takers from U.S. medical schools, the vast majority had recently completed their second year of medical school; thus, the number of weeks of full-time study was dependent to some degree on when they had completed their coursework. Analysis indicated that the examinees had studied for an average of 5.8 weeks and for 53 hours per week. Scores were positively correlated with weeks of full-time study. However, Step 1 performance showed little relationship to the number of hours of study per week.

Regression Analysis of Factors Influencing Step 1 Scores. A series of regression analyses were run to examine the relationships between Step 1 performance and exam-preparation strategies. The first analysis, the "full model," included all four sets of predictors described previously: pre-matriculation characteristics, study time, preparation methods, and coaching course participation. Then, for each set, two regression equations were estimated: the first included just the predictors in the set, and the second included all predictors except those in the target set. The R^2 for the first equation provides an upper bound on the variance explained by the set; the difference between the R^2 for the full model and the R^2 for the second model provides a lower bound (the proportion of variance uniquely predicted by the set). The full model explained 33.6% of the variance in Step 1 scores. Table 2 provides the upper and lower bounds on the R^2 predicted by each set.

51

Table 1. Mean USMLE Step 1 Scores by Test-preparation Approach Used

Approach	Approach Used	Approach Not Used	*p* Value*
General instructions booklet			
Mean score	222.91	219.90	.009
No. students	851	366	
SD	18.69	18.43	
Lecture notes			
Mean score	222.64	221.61	.349
No. students	468	749	
SD	18.69	18.84	
Note-taking service			
Mean score	224.12	221.88	.327
No. students	67	1150	
SD	18.70	18.62	
Textbooks			
Mean score	224.90	219.70	<.001
No. students	540	677	
SD	18.28	18.64	
Course syllabi			
Mean score	226.66	220.80	<.001
No. students	250	967	
SD	18.29	18.56	
Commercial books about USMLE			
Mean score	222.22	213.28	.010
No. students	1188	29	
SD	18.43	23.57	
School materials about USMLE			
Mean score	221.43	221.43	.060
No. students	304	913	
SD	18.86	18.52	
Study in groups			
Mean score	221.50	222.18	.576
No. students	309	908	
SD	18.48	18.68	
Coaching course			
Mean score	217.38	223.38	<.001
No. students	279	938	
SD	19.83	18.05	

* From independent-samples t-test.

Results indicated that pre-matriculation characteristics accounted for nearly all of the explained variance. Study time explained a small amount of variance, with both weeks of study and hours of study per week having a small positive influence on scores. However, neither coaching courses nor preparation methods had a significant influence on scores. Further, the term measuring the interaction between coaching course enrollment and MCAT scores was not significant.

In order to focus on preparation methods in more detail, we examined each individual method separately, controlling for the pre-matriculation characteristics. The results showed that use of textbooks had a significant effect on Step 1 scores, although the effect was fairly small (1.9 points). The other preparation methods did not significantly affect scores.

Discussion

Studies of the possible effects of coaching courses have been reported for several post-secondary exams, but have been notably

Table 2. Regression Results Concerning the Relationships between USMLE Step 1 Performance and Four Sets of Predictors*

Variable Set	R^2	Unique R^2
Full model	.335	—
Pre-matriculation (MCAT, Science GPA)	.324	.282*
Study time	.008	.004*
Study methods	.035	.004
Coaching courses	.018	.001

*$p < .05$.

absent for the USMLE; the only studies we could locate were undertaken in the 1970s at single medical schools for the NBME Part I exam. In the first of these studies, Scott et al.[6] found significantly higher scores in coached examinees in only one of the three years studied. They found that coaching offered greater benefit to students with lower basic science GPAs (first two years of medical school coursework) than it did to students with higher basic science GPAs. Students were surveyed as part of that study, and the vast majority of students thought the course had been beneficial. Both the authors and the students surveyed cited the relevance of the course content to Part I and the organization of the material as the most valuable features of the course. In contrast, Lewis and Kuske[7] reported that after controlling for the examinees' basic science GPAs, commercial review courses had no detectable effect on scores.

In the present study, the examinees had used several different strategies while preparing for the USMLE Step 1. By far the most common approach was to use commercially prepared study guides. These had been used by 98% of the survey respondents, indeed by more respondents than had used the traditional textbooks they are generally required to buy. The use of commercial study guides also eclipsed the use of materials prepared or approved by professional medical school educators, and eclipsed use of the USMLE publication designed specifically for Step 1 preparation. Ironically, the results obtained indicate that examinees may benefit by using standard texts, many of which they have already purchased. There was little or no evidence of achievement of higher scores as a consequence of using commercially prepared material, controlling for pre-matriculation characteristics and other study methods used.

Perhaps the most interesting finding in this analysis is the limited impact of coaching courses on scores. Several caveats must accompany this finding. First, students who enroll in coaching courses are self-selected, in some instances because they are concerned about their readiness to take Step 1. This self-selected group may disproportionately include students in academic trouble at their medical schools and those who have been warned that they are in danger of failing based on tests given in medical school. Second, time-intensive courses may compete with other preparation methods that are more effective or time-efficient. Third, our study lumped together all coaching courses, which vary in length, intensity, and teaching methods. With these provisos, our findings suggest that participation in coaching courses appears to have little effect on scores when controlling for educational antecedents, time of study, and other preparation methods.

Last, the reduction in sample size due to survey return rate and incomplete data was problematic. Examinees included in the study had slightly higher Step 1 scores (220.7 versus 216.1). Thus, there is some evidence to support the intuitive reasonableness of selection bias.

Correspondence: Raj A. Thadani, MA., NBME, 3750 Market Street, Philadelphia, PA 19104; e-mail: (rthadani@mail.nbme.org).

References

1. Powers D. Who benefits from preparing for a "coachable" admissions test? J Educ Meas. 1987;25:247–62.
2. Ornstein AC. Coaching, testing and college admission scores. NASSP Bul. October 1993;12–9.
3. Jones RF. The effect of commercial coaching courses on performance on the MCAT. J Med Educ. 1986;61:273–84.
4. Koenig JA, Leger KF. A comparison of retest performances and test preparation methods for MCAT examinees grouped by gender and race-ethnicity. Acad Med. 1997;72(10 suppl):S100–S102.
5. Swanson DB, Case SM. A preliminary study of the validity of the new Medical College Admission Test for predicting performance on USMLE Step 1 and Step 2. Acad Med. 1996;71(1 suppl):S25–S27.
6. Scott LK, Scott CW, Palmisano PA, Cunningham RD, Nass JC, Brown S. The effects of commercial coaching for the NBME Part I examination. J Med Educ. 1980;55:733–42.
7. Lewis LA, Kuske TT. Commercial national board review programs: a case study at the Medical College of Georgia. JAMA. 1978;240:754–5.

53

Effectiveness of Telehealth for Teaching Specialized Hand-assessment Techniques to Physical Therapists

WENDY BARDEN, HOWARD M. CLARKE, NANCY L. YOUNG, NANCY McKEE, and GLENN REGEHR

Health care reform has changed the focus of patient care from primarily inpatient to an increased emphasis on outpatient services. The reductions in hospital beds, staff complements, and lengths of inpatient stays have led to an increased need for early referral of patients to health care professionals in the community. Unfortunately, a corresponding adjustment of outpatient resources has not occurred, resulting in an imbalance of resources and making accessibility to the appropriate services in the community extremely difficult.[1] This is especially true for outlying communities, since metropolitan centers have disproportionately large numbers of health care specialists.[2,3] In North America, with its vast geographic areas, travel between the metropolitan centers and the rural communities is often problematic, creating difficulties for patients needing care and for rural practitioners, who experience a feeling of professional isolation.[4]

Telehealth technology may provide an economically feasible solution to these concerns. Telehealth has been defined as the utilization of telecommunications technology to provide health care services and medical information over distance.[5] Telehealth has the potential to improve services to rural communities by providing not only direct telemediated access to clinical specialists for patients, but also the opportunity for the efficient training of rural professionals in the necessary specialty care.[6]

A broad range of medical specialties has demonstrated the capabilities of telehealth to assess patients in remote areas.[7] Much of this research, however, has focused on domains in which visual[8,9] and/or auditory[1] information is sufficient for accurate assessment. It is less clear, however, whether telehealth assessment is equally effective for specialties where tactile interaction between the patient and health care professional is considered critical. For these situations the health care professional at the distant site must be the "consultant's hands."

There is a parallel in using telehealth for the purposes of clinical education. That is, telehealth may be effective for teaching knowledge-based topics, but many health profession domains, such as physical therapy assessment skills, have tactile components that require measurement and analysis. Training for these types of skills may challenge the application of telehealth beyond its current capabilities. The purpose of this research, therefore, was to determine the effectiveness of telehealth for teaching specialized assessment skills requiring "hands on" interaction with patients.

Method

Participants. In 1999, a total of 42 physical therapists from two Northern Ontario cities agreed to participate. They were stratified by city and were systematically allocated to one of three interventions to ensure that the groups were balanced according to age, graduation year, type of educational format utilized at the university where the participants trained, prior hand therapy experience, prior telehealth experience, and type of current clinical practice.

Interventions. Three educational formats were used to teach five hand-assessment skills: volumetrics of the hand; total active movement of the index finger; joint mobilization of the proximal interphalangeal joint of the long finger; grip strength; and two-point discrimination of the ulnar nerve. The three interventions were self study (SS); direct face-to-face teaching (DT); and telehealth teaching (TT). The same information was provided to the therapists in each of the three formats. However, the manners in which this information was transmitted differed across the three formats.

The therapists assigned to the SS group were provided with a package containing written information thar they were able to review over a three-day period. This material outlined how to correctly perform each hand-assessment skill based on the guidelines established by the American Society of Hand Therapists. There were approximately three pages of information per skill, including history, indications, contraindications, technique, and diagrams demonstrating performance of the skill. When given the documentation, these therapists were given instructions to learn the material independently in the same manner as they would normally.

The DT session involved approximately 3.5 hours of direct contact with the instructor and was organized such that the instructor taught each skill for 15 minutes, providing the relevant information as described above and demonstrating the skill using a standardized patient. Immediately following the teaching and demonstration of each skill, the therapists practiced in pairs for approximately 30 minutes using each other as the "patient," with the expectation that when they were not interacting with the instructor they would exchange ideas to solve problems and to perfect their performances. During the 30-minute practice period each pair also received five minutes of direct contact and interactive feedback from the instructor.

The TT session was identical in format and timing to the DT session. To ensure similarity of presentation, the primary investigator of the study was present at both teaching sessions. The primary difference between the DT and TT groups was that the participants were located together at one local telehealth site and the same instructor from the DT group was located at a second local telehealth site that was physically removed from the first. Participant–instructor interaction was therefore mediated using an intracity link between two facilities that housed compatible videoconferencing equipment, thus eliminating the possibility of the direct "hands-on" contact with the instructor during the interactive feedback components of the session.

Evaluation Instruments. A modified objective structural clinical examination format was used for both a pre-test evaluation and a post-test evaluation. Each participant performed the five skills consecutively on a single standardized patient, taking up to five minutes per skill. All five skills were evaluated by the same examiner (a content expert who was blinded to the intervention condition), with a separate mark given for each skill. Two evaluation instruments were used for each skill. First, a five-point global rating scale with four domains—knowledge of the technique, the ability to perform the technique, instrument handling, and organizational skills—was used to assess the underlying characteristics of performance. Anchors were provided for points 1 (poor, unable to perform), 3 (adequate), and 5 (excellent performance). The global score for each skill was calculated as the average of the scores for the four separate domains. Pilot work on this global scoring technique confirmed inter-rater reliability ($ICC_{2,1}$ = 0.78–0.91 for the five skills) and construct validity (with skill level—novice versus intermediate versus expert—accounting for 20–67% of the variations in scores for the five skills). As a second measure of performance, the examiner completed a binary question addressing competency for each skill.

54

Procedure. The research was conducted over two five-day periods one month apart in each of two Northern Ontario cities. For each city, the participating therapists individually attended a pre-test. These were scheduled for 30 minutes, and all were completed over a two-day period. Following each participant's pre-test, the participant was given instructions relevant to his or her teaching intervention. Those in the SS group were given the manual with appropriate instructions and given a time for their post-test session. Participants in the TT and DT groups were told when and where to arrive for the instructional session and were given a time for their post-test session. All participants were asked to avoid discussing the nature of the test with other participants prior to completion of the pre-test period, and participants from each group were asked not to discuss the training material across groups in order to avoid contamination of the experiment. At each city the TT session was held in the morning and the DT session was held in the afternoon of the third day. The post-test was conducted over the last two days, with the relative time of the post-test for each participant as close as possible to the relative pre-test time to ensure almost-identical delays between pre- and post-tests for all participants.

Results

Performance Scores. The summary statistics for the performance scores for all five skills are presented in Table 1. It is clear that the DT and TT groups approached excellent performance on all five skills after the intervention, whereas the SS group demonstrated only adequate performance on three of the five skills and poor performance for the two remaining skills. For all five skills, the interaction terms from the two-way ANOVAs suggest significant differences in the amounts of learning among the groups ($F_{2,19}$ values ranged from 4.98 to 26.65, for all analyses, $p < .01$). The subsequent one-way ANOVA comparing the three groups on the pre-test showed no effect for any of the five skills ($F_{2,19}$ values all less than 1.00, ns), suggesting that all three groups started at the same skill level. However, the one-way ANOVA comparing the three groups on the post-test showed powerful, significant effects of the group ($F_{2,19}$ values ranged from 9.96 to 35.06, for all analyses $p < .01$), suggesting differences in abilities among the three groups after the intervention. A series of post-hoc Tukey tests demonstrated no significant difference between the DT and TT groups but a significant difference between the SS group and both the DT and TT groups, suggesting that the members of the DT and TT groups learned equally well, and learned better than did those in the SS group. Finally, given the lower post-test scores for the SS group, a series of paired t-tests was performed on the SS group results to determine whether the SS group was a worthwhile intervention. For only three of the five hand assessment skills was there a significant difference in the pre-test and post-test performance ($p < 0.05$), and the difference scores for these those skills are relatively small, raising the question of whether the difference is clinically significant (see Table 1).

Competency Scores. A similar pattern of results occurred with the competency scores. The pre-test and post-test percent competency scores are presented in Table 2. Chi-square analyses of the pre-test competency assessments showed no significant differences between groups on any of the five skills (χ^2_2 ranged from 0.20 to 1.03, ns, with two being incalculable because all participants were evaluated as not competent), suggesting that individuals from each group were equally likely to be competent. By contrast, chi-square analyses of the post-test results revealed significant effects of group for all five skills (χ^2_2 ranged from 8.42 to 24.21, for all analyses, $p < .01$). A series of subsequent chi-square analyses comparing the methods by pairs again showed no significant difference between the DT and TT groups, but significant differences between the SS and TT groups and significant differences between the SS and DT groups for all five skills. Finally, a series of subsequent McNemar's tests was performed on the SS group competency results to determine whether this intervention was able to change the competency levels of subjects. For all five skills there was no significant change in competency levels for the SS group.

Discussion

The primary purpose of this study was to determine whether telehealth could be utilized to effectively teach specialized assessment skills to physical therapists. This study has demonstrated that telehealth may be used in this capacity.

The five hand-assessment and treatment techniques that were selected for this study all possessed components that would challenge the transmission capabilities of telehealth. Three of the skills—volumetrics, total active movement, and grip strength—required the participants to use primarily visual learning skills. All elements of these three skills are easily learned by watching a demonstration or studying written material. Therefore, it was of no great surprise that these three skills were successfully taught via telehealth. What was somewhat surprising were the low pre-test performance and competency scores for the grip-strength technique, since this skill is simple and frequently used in many areas of physical therapy. However, the low scores are easily explained by the strict guidelines set by the American Society of Hand Therapists that were used during the evaluation of the participating therapists.

The two remaining skills, joint mobilization and two-point discrimination, are not strictly visual but are skills that require tactile input. Initially, there was a concern on how transmission of tactile feelings could be transmitted via telehealth. With clear, concise instructions and appropriate camera placement it was demonstrated that these two skills could be learned.

In this study, it was demonstrated that telehealth teaching, when compared with the conventional teaching model of direct face-to-face teaching, resulted in no statistically significant difference between the performance scores for any of the five skills taught. However, when compared with self-study, there were statistically significant differences in the performance scores, suggesting that the telehealth group learned more. Both of these results suggest that telehealth may be used as effectively as the conventional method and more effectively than self-study to teach these five assessment skills.

In examining the competency scores for the telehealth and the direct, face-to-face groups there was once again no statistically significant difference between the groups at baseline. Differences in the competency levels were determined, however, after the educational intervention, indicating that the groups had become more competent in all five skills. When the telehealth group was compared with the self-study group, there were statistically significant differences between the groups' competency scores for all five skills.

The results of this study must be interpreted with some caution. We did not, for example, ask the participants in the self-study group what they had done to prepare for the post-test. Thus, although we asked them to do what they would normally do if a patient being referred required that technique, we do not know what the participants in the SS group actually did or the length of time they might have spent preparing relative to the time spent in the formal intervention groups. It is unlikely that they spontaneously practiced, and even more unlikely that they sought external feedback for their practice, two components of the formal training programs that were likely very important. Further, we did not ask them what they would normally do in these circumstances, so without further study we cannot say whether the SS group's performance is representative of normal practice.

Similarly, we do not know the extent of contamination between the groups. Although the participants were specifically asked not

55

TABLE 1. Mean Pre- and Post-test Scores of 42 Physical Therapists (in Three Groups) on Five Skills, University of Toronto, 1999*

Skill	Group	Pre-test Mean (SD)	Post-test Mean (SD)	Difference
Volumetrics	Self study	2.23 (1.23)	3.19 (1.25)	0.95 (1.61)‡
	Direct teaching	2.17 (1.42)	4.63 (0.46)†	2.46 (1.36)
	Telehealth teaching	2.23 (1.36)	4.50 (0.40)†	2.27 (1.13)
Total active movement	Self study	1.45 (0.40)	2.23 (1.08)	0.78 (1.21)‡
	Direct teaching	1.38 (0.51)	4.33 (0.75)†	2.94 (0.65)
	Telehealth teaching	1.58 (0.68)	4.65 (0.59)†	3.08 (0.87)
Joint mobilization	Self study	2.98 (1.69)	3.66 (1.33)	0.67 (1.29)§
	Direct teaching	2.62 (1.41)	4.92 (0.28)†	2.31 (1.46)
	Telehealth teaching	2.69 (1.60)	4.85 (0.42)†	2.15 (0.51)
Grip strength	Self study	2.92 (1.22)	3.06 (0.84)	0.14 (1.52)§
	Direct teaching	3.06 (1.21)	4.69 (0.23)†	1.63 (1.14)
	Telehealth teaching	3.13 (1.24)	4.35 (1.13)†	1.21 (1.23)
Two-point discrimination	Self study	1.11 (0.26)	1.84 (1.03)	0.73 (1.05)‡
	Direct teaching	1.13 (0.22)	4.10 (0.79)†	2.96 (0.83)
	Telehealth teaching	1.12 (0.35)	4.04 (0.89)†	2.92 (0.93)

*Scores were on a global rating scale (1 = poor, unable to perform; 3 = adequate knowledge; 5 = excellent performance). See text for description of the three groups and the pre- and post-tests.
†Significantly different from self study on post-test by post-hoc Tukey test ($p < 0.05$).
‡Significant improvement from pre- to post-test using paired t-test ($p < 0.05$).
§No significant improvement from pre- to post-test using paired t-test.

to interact between groups, we did not subsequently determine the extent to which they had followed these instructions. This might limit the validity of the findings, although it is worth noting that the group that had more motivation to violate this injunction to speak to others continued to have lower scores.

Despite these potential limitations, the current study gives us great hope for the use of the telehealth medium for teaching not only technical information but also technical skills. Establishing telehealth as an effective teaching tool provides a method of continuing education to community health care professionals who need to perform these types of technical skills. Therefore, all professionals (nurses, therapists, doctors) would benefit from this technology, allowing increasingly early referral of complex cases to the community for ongoing rehabilitation. If telehealth is utilized to transmit and teach the required information, continuity of specialized care will be maintained with support provided to the community practitioner. Perhaps teaching of all assessment skills will not be possible, but telehealth will continue to provide a rich communication link between the acute care facilities and the community.

TABLE 2. Percentages of 42 Physical Therapists Identified as "Competent" in Each of Three Groups, on Five Skills, University of Toronto, 1999*

Skill	Group	% after Pre-test	% after Post-test	Change
Volumetrics	Self study	37.5	56.3	18.8‡
	Direct teaching	30.8	100.0†	69.2
	Telehealth teaching	30.8	100.0†	69.2
Total active movement	Self study	00.0	12.5	12.5‡
	Direct teaching	00.0	76.9†	76.9
	Telehealth teaching	00.0	92.3†	92.3
Joint mobilization	Self study	56.3	75.0	18.8‡
	Direct teaching	53.8	100.0†	46.2
	Telehealth teaching	38.5	100.0†	61.5
Grip strength	Self study	68.8	50.0	−18.8‡
	Direct teaching	61.5	100.0†	38.5
	Telehealth teaching	61.5	84.6†	23.1
Two-point discrimination	Self study	00.0	6.3	6.3‡
	Direct teaching	00.0	76.9†	76.9
	Telehealth teaching	00.0	76.9†	76.9

*See text for description of groups and the pre- and post-tests.
†Significantly different from self study on post-test by chi-square ($p < .05$).
‡No significant improvement from pre- to post-test using McNemar test

56

This research was funded through grants from the G. H. Wood Foundation and the Research Institute at the Hospital for Sick Children, Toronto, Ontario, Canada. This research was also supported by the Division of Plastic Surgery, Department of Rehabilitation Services, and the Telehealth Programme at the Hospital for Sick Children, Toronto, and the Faculty of Medicine Centre for Research in Education at The University Health Network, Toronto, Ontario, Canada.

Correspondence: Wendy Barden, Department of Rehabilitation Services, Hospital for Sick Children, 555 University Avenue, Toronto, Ontario, Canada M5G 1X8; e-mail: ⟨wendy.barden@sickkids.on.ca⟩.

References

1. Cheung ST, Davies RF, Smith K, et al. The Ottawa Telehealth Project. Telemedicine J. 1998;4:259–66.
2. Jones E. Telemedicine brings care to far-flung provinces. Telemedicine and Telehealth Networks. 1996;May(5):5–6.
3. Dunn EV, Higgins CA. Telemedicine in Canada: an overview. Dimensions. 1984; July(7):16–8.
4. Brauer GW. Telehealth: the delayed revolution in health care. Medical Progress through Technology. 1992;18:151–63.
5. Klotz J. Telemedicine is here to stay. J Cutaneous Medicine and Surgery. 1998;2: 224–5.
6. Jennett PA, Hall WG, Morin, JE, et al. Evaluation of a distance consulting service based on interactive video and integrated computerized technology. J Telemedicine and Telecare. 1995;1:69–78.
7. Picot J. Telemedicine and Telehealth in Canada. Forty years of change in the use of information and communications technologies in a publicly administered health care system. Telemedicine J. 1998;4:199–205.
8. Scerri GV, Vassallo DJ. Initial plastic surgery experience with the first telemedicine links for the British Forces. Br J Plast Surg. 1999;52:294–8.
9. Nitzkin JL, Zhu N, Mariet R. Reliability of telemedicine examination. Telemedicine J. 1997;3:141–57.

57

A Controlled Trial of an Interactive, Web-based Virtual Reality Program for Teaching Physical Diagnosis Skills to Medical Students

JULIE A. GRUNDMAN, ROBERT S. WIGTON, and DEVIN NICKOL

Multimedia instruction offers many potential advantages over traditional methods of instruction. Multimedia programs can interact with the learner, use graphic images, sound, and video, and keep track of progress. Students complete programs at their own pace while accessing material both at school and at home. Multimedia instruction can provide an interactive alternative to lectures and textbooks by quizzing the student over concepts as they are presented and requiring that the student think about the material before proceeding.

While several studies have found that multimedia instruction can be more efficient by reducing instructor and classroom time, few have been able to show an increase in learning when compared with traditional methods of instruction.[1-6] Santer and colleagues compared a multimedia textbook with a lecture presentation on the same material and found an increase in the post-test scores of the multimedia group, but no difference when they compared the multimedia group with a group using a printed textbook.[1] Studies comparing multimedia and traditional approaches to learning in the areas of psychology and computer science instruction suggest an improvement in students' performances using rhe multimedia versions.[7,8] Thus, there is a need for well-designed studies to determine whether multimedia instruction more effectively facilitates students' learning—including medical students' learning—than do traditional methods.

Multimedia instruction is particularly well suited to help students learn physical diagnosis. Sound, pictures, and movies augment the learning of examination skills and diagnosis findings by allowing students to hear heart and lung sounds, watch videos of physical examination procedures, and see more pictures of pathologic findings than can be included in a textbook or lecture. These visual and audio aids should increase students' recognition of these findings when encountered in patients.

We wished to test whether a Web-based multimedia program using interactive learning and virtual reality would be more efficient and effective than traditional print-based self-study by medical students. To accomplish this, we designed a course on physical examination of the eye and ear. Using this material, we conducted a controlled study of first-year medical students to determine whether having students use a multimedia version of the course resulted in a change in the time spent with the material and an increase in knowledge gained when compared with having students study a printed version of the same material.

Method

Participants. All 126 first-year medical students at the University of Nebraska Medical Center College of Medicine were invited to participate in this study in 1999; of these, 121 volunteered. We obtained Institutional Review Board approval and participants gave informed consent.

Multimedia Course. We designed two courses, one about the eye and the other about the ear, to help first-year medical students learn physical diagnosis skills and tests. To minimize contamination, the subject matter chosen represented important clinical skills that ordinarily would not be presented at this time and was not part of the regular first-year curriculum. Besides otoscopy and funduscopy, topics included the interpretation of audiograms and tympanograms, as well as recognition of acute otitis media, serous otitis media, papilledema, glaucoma, and diabetic retinopathy. Some of the topics were chosen because we predicted they could be more effectively taught multimedia presentations. We estimate the course took 300 hours to create at a cost of $300 in materials and $3,000 in student labor.

After developing the courses, we created both a multimedia and a printed version of the course materials. The multimedia version emphasized interactivity and included many pictures and QuickTime virtual reality (QTVR) movies of the funduscopic examinations and otoscopy. These movies showed normal and pathologic views of the retina and tympanic membrane as they would be seen through an ophthalmoscope or otoscope. Students were asked to scan around the entire picture searching for pathology, just as they would when examining actual patients. In addition, the multimedia version led students on a defined path through the program by presenting a concept on each page and then requiring that the student respond to a question about the concept before proceeding. Correct answers allowed the student to proceed while students giving incorrect answers were provided with immediate feedback before continuing. The multimedia version differed from the printed version in that it included interactive questions and contained more color pictures.

Study Design. The students took a pre-test before the courses and a post-test afterwards. The pre-test consisted of 20 multiple-choice questions about physical examination skills and findings related to the eye and ear. The class had already been divided into 12 small study secrions. In order to provide two groups matched with regard to their knowledge of the subject matter, we divided the existing study sections into two groups matched on average pre-test score. Group A (n = 60) used the printed manual for the ear material and the multimedia program for the eye material. Group B (n = 61) did the reverse.

Students could access the materials one week before the post-test. We used three methods to track the time spent. The computer logged access and time spent with the mulrimedia material. To use the printed version, students had to check it in and out of the library reference desk, creating a log of the time spent. In addition, we asked the students to keep their own records of the time spent. This was reported on a post-study questionnaire. As an incentive, the student with the top score in each of the two groups was awarded a $25 cash prize.

The post-test was given in two sections. The first portion consisted of ten multiple-choice questions presented on the computer and included questions regarding virtual-reality simulations of fundoscopy and otoscopy. The second portion of the post-test, given on the following day, consisted of 40 multiple-choice questions, 20 on the eye and 20 on the ear. These questions included general fact-based questions, images, and case studies. All questions and images were different from those used in either the written or the multimedia version.

Evaluation and Analysis. The level of the students' acceptance of the program was evaluated with a written survey. The students were asked which instructional method they preferred, how much time they had spent, and how they rated their learning using each method.

58

TABLE 1. Post-test Scores and Time Spent by Instruction Method, University of Nebraska College of Medicine, 1999*

	Multimedia Method					Printed-version Method				
		Questions correct					Questions correct			
Skill	Group	Mean	%	95% CI†	Mean No. Min.	Group	Mean	%	95% CI†	Mean No. Min.
Eye	A	15.9	64	15.0–16.8	39.9	B	13.4	54	12.6–14.2	31.0
Ear	B	16.1	64	15.3–16.9	48.6	A	14.5	58	13.6–15.4	36.0

*A total of 121 first-year medical students (60 in Group A, 61 in Group B) took the post-test to examine their skills in topics related to physical diagnosis skills and tests concerning the human eye and ear. All comparisons between multimedia and printed version are significant at $p < .02$ (MANOVA).

†Confidence interval.

Post-test scores were analyzed using multiple regression and ANCOVA, conrrolling for time spent on the multimedia or printed versions, section, pre-test scores, scores on the Medical College Admission Test (MCAT), and college grade-point average (GPA).

Results

To ensure that the two groups were equivalent. we compared them with regard to their members' MCAT sections' scores, undergraduate GPAs, and mean pre-test scores. The pre-test scores were similar, with group A correctly answering 8.65 of 20 (43.3%) and group B correctly answering 8.52 of 20 (42.6%). There was no significant difference between the groups' mean Verbal Reasoning MCAT scores, Physical Sciences MCAT scores, Biological Science MCAT scores, total undergraduate GPAs, and pre-test scores. Reliability for the paper-based portion of the post-test was 0.69 using the Kuder Richardson formula 20.

The students who used the multimedia version of the eye or ear course scored higher on the respective post-tests than did the students who used the printed version, when compared using univariate analysis (see Table 1). This higher achievement in the multimedia version persisted when pre-test scores, MCAT scores, and undergraduate GPAs were controlled for by entering those variables into multivariate analysis. Multiple regression analysis showed that both use of the multimedia version and the time spent were predictors of the eye course's post-test score ($r^2 = 0.24$). Only the use of the multimedia version (and not time spent) predicted the ear course's post-test score ($r^2 = 0.11$). Interestingly, the students who used the multimedia version did not score higher on the subset of questions dealing specifically with the virtual-reality simulations or on the computer-based subset.

Students using the multimedia version spent more time on the material than did those using the printed version (see Table 1). However, an increase in time could have resulted in an increase in the post-test score. With this possibility in mind, we found that when time and pre-test score were controlled for using ANCOVA, students using the multimedia version still performed better than did those using the printed manual ($p < .001$). The only correlation between time and post-test score occurred with the multimedia version of the eye information ($r^2 = 0.61$, $p < .0001$).

The results of the written survey showed that 78% of the students preferred the multimedia version to the printed version and were interested in using similar programs for other areas of physical diagnosis. Most students indicated that their learning had been more effective using the multimedia version, and stated that they had enjoyed the virtual-reality movies and interactivity. The students did not report any difficulry in accessing either the computer or the written version of the program.

Discussion

As described above, we found that students using an interactive multimedia program improved more than did those using a printed manual with the same content. The students spent more time using the multimedia version but also improved more, given the time spent. These findings indicate that the multimedia program was more effective.

Which aspects of the program led to this increase in post-test score? Results showed that the increased time accounted for some, but not all, of the gain. The gain in knowledge could have been due to the increased time that students spent on the multimedia version, but this did not appear to be the case, since even after taking time into account, the multimedia program still showed a greater improvement. The virtual-reality movies may have contributed, but our results showed improvement in all parts of the post-test, not just those dealing with virtual reality. The interactive dialog probably played a large role in the program's effectiveness by encouraging the students to work through problems, inducing them to take more time on particular tasks and probably to give more attention to the material. The computer itself could have affected the results, but there was no indication that this was the case, since the multimedia group had a higher post-test score but not a higher score on only the computer-based questions.

There are several limitations to the generalization of these results. We have not yet measured the long-term retention of the information, or whether the students who used the multimedia version perform better when examining patients. The latter activity may be best assessed using a performance-based assessment tool, such as an objective structured clinical examination. The multimedia version had more illustrations, but the results suggest this was not a major cause of the differences in achievement, since the scores were not better on the computer part or on the few questions that involved pictures or feature recognition. The ability to include more pictures is an inherent advantage of multimedia programs over textbooks, where the cost of color pictures generally precludes their inclusion. Finally, we studied first-year students at only one medical school, so the generalizability of these results in other settings and other courses should be confirmed.

In our study, we selected two specific content areas that we thought were well suited to multimedia. But other areas of physical diagnosis may show a similar benefit when using multimedia as a learning resource, such as having the ability to listen to heart and lung sounds. Studies have already shown this to be of benefit in teaching cardiac ausculatory skills.[1]

Is multimedia instruction more efficient in the medical school setting? Multimedia programs can be reused and offer flexible scheduling, but complex programs may take considerably more time to develop. Lyon found that a multimedia program reduced instructor time with no loss of student achievement.[6] In our study, the students spent more time with the multimedia version, and the time spent resulted in greater achievement, but this version required greater development time. Our experience suggests that whether a multimedia program results in gains in efficacy may depend on the nature of the subject and the learning mode it replaces.

59

Conclusion

Our study suggests that multimedia learning, incorporating interactivity and virtual reality, is more effective than traditional approaches to teaching the eye and ear sections of the physical examination. Learning was enhanced in all areas, not just those dealing with virtual reality or the multimedia. The implications are that more multimedia courses in physical diagnosis techniques should be developed and evaluated. Further study is needed to determine what aspects of multimedia learning are most effective and how well the results found here will apply to other areas of the curriculum.

Correspondence: Robert S. Wigton, MS, MD, Section of General Internal Medicine, University of Nebraska Medical Center, 984285 Nebraska Medical Center, Omaha, NE 68198-4285; e-mail: (Wigton@unmc.edu).

References

1. Santer DM, Michaelsen VE, Erkonen WE, et al. A comparison of educational interventions. Arch Pediatr Adolesc Med. 1995;149:297–302.
2. Mangione S, Nieman LZ, Greenspon LW, Margulies H. A comparison of computer-assisted instruction and small-group teaching of cardiac auscultation to medical students. Med Educ. 1991;25:389–95.
3. Criley JM, Criley D, Zalace C. Multimedia instruction of cardiac auscultation. Trans Am Clin and Clim Assoc. 1996;108:271–84.
4. Mangione S, Nieman LZ, Gracely EJ. Comparison of computer-based learning and seminar teaching of pulmonary auscultation to first-year medical students. Acad Med. 1992;67(10 suppl):S63–S65.
5. Keane DR, Norman GR, Vickers J. The inadequacy of recent research on computer-assisted instruction. Acad Med. 1991;66:444–8.
6. Lyon HC, Healy JC, Bell JR, et al. PlanAlyzer, an interactive computer-assisted program to teach clinical problem solving in diagnosing anemia and coronary artery disease. Acad Med. 1992;67:821–8.
7. Erwin TD, Rieppi R. Comparing multimedia and traditional approaches in undergraduate psychology classes. Teaching of Psychology. 1999;26:5861.
8. Crosby ME, Stelovsky J. From multimedia instruction to multimedia evaluation. J Educ Multimedia and Hypermedia. 1995;4:147–62.

60

Evaluation of a CME Problem-based Learning Internet Discussion

JOAN M. SARGEANT, R. ALLAN PURDY, MICHAEL J. ALLEN, SHAILESH NADKARNI, LINDA WATTON, and PEARL O'BRIEN

Research into the impact of continuing medical education (CME) demonstrates that effective interventions include practice-enabling or reinforcing strategies, sequential activities, and/or a high degree of interaction among participants.[1,2] Problem-based learning (PBL), a strategy used in CME, engages participants in small-group interactive learning, creating a context that reflects the practice setting by presenting actual cases as problems to be solved.[3] PBL specifically has been shown to be an effective learning strategy in CME.[4,5]

Traditionally, PBL participants have been required to be in the same place at the same time, but now the Internet enables interpersonal interaction that is independent of time and place. Using asynchronous (delayed) interaction via a bulletin board, learners in different locations can participate in on-line discussions at times convenient for them.[6] For physicians this removes barriers (e.g., geographic location, practice responsibilities) to participating in conventional CME programs and interacting with fellow learners.[7]

Although the Internet provides many opportunities for medical education,[8-10] a recent search of the medical literature revealed few studies of its use for interpersonal interaction in medical education,[9-16] and only one study of on-line PBL. In that study, Chan[17] attempted to determine the effectiveness of an on-line PBL program in a randomized controlled trial of 23 physicians. Group process, however, was not the focus of study, and the number of messages was small (35 over two months).

Barrows advocates that successful PBL requires facilitators to perform four functions: navigating (guiding the group through the activities), facilitating (maintaining a constructive group process), questioning (using questions to deepen understanding), and diagnosing (monitoring learners' progress).[18] Berge and Collins suggest facilitator roles for general on-line discussions, which include pedagogic (ensuring the educational task is accomplished), social (creating a friendly environment), managerial (administrating organizational elements), and technical (ensuring comfort with the technology) roles.[19]

The purposes of the present study of an Internet (on-line) CME PBL discussion were (1) to describe the roles of facilitators, both on-line and off-line, that enable on-line discussion; (2) to determine factors that influence learners' participation in the on-line discussion; and (3) to determine learners' satisfaction with the on-line discussion. The study was a process evaluation, which documents and assesses the implementation of a program's activities to guide further program planning.[20]

Method

Family practitioners in Nova Scotia comprised the target population, but other physicians could register. We recruited participants by advertising the program locally and demonstrating it at a provincial CME event. The intervention, carried out in 1999, was an on-line case-based learning module on medication-induced headache (MIH) developed by a neurologist for a conventional PBL CME workshop and modified for Internet presentation. We chose this program because the original workshop was successful[5] and the neurologist is an expert PBL facilitator interested in Internet learning.

We used Web-CT[21] educational courseware for the module. Besides the case-based discussion in the bulletin board, the module included a "lecture," a quiz, a glossary, and references. We encouraged learners to review the lecture before joining the discussion. To meet the College of Family Physicians of Canada accreditation criteria for on-line CME programs; i.e., that the program be available for a defined time period and provide the opportunity for physician interaction,[22] the module was available for one month and participants were required to post at least one message in the bulletin board.

Using Berge and Collins' facilitator roles,[19] we outlined two general roles for the facilitators of the on-line PBL discussion. These were (1) the pedagogic, or content, role, assumed by the neurologist or content facilitator, and (2) a combined social (creating a supportive environment), managerial, and technical role, assumed by two educators. A graduate student familiar with Web-CT also provided technical support.

We collected data using the Web-CT electronic activity record, the program evaluation questionnaire, facilitators' records of on-line and off-line activities, bulletin board discussion transcript, a log of technical problems, and interviews with registrants who did not participate. The questionnaire was designed to evaluate all components of the on-line program and consisted of 51 closed-ended and seven open-ended questions. For this study, we used the nine closed-ended and one open-ended questions that addressed the case and bulletin-board discussion, and three closed-ended and one open-ended question that addressed the general usefulness of the module. Participants completed the questionnaire electronically or on paper.

Evaluation questionnaires received electronically were automatically entered into the Web-CT database, which computes descriptive statistics. We manually entered evaluations received by paper. For the bulletin board discussion transcript, we used content analysis to categorize data and identify themes.[23]

Results

The 31 registrants were 28 family physicians, two family medicine residents, and one neurologist. The electronic activity record showed that 12 registrants did not participate. Of these, three did not log into the program, and nine accessed the home page only. We attempted to contact these 12 and received responses from four. Two were unable to log on and had not contacted the "help line." One reported personal computer failure and another had become "too busy." Of the 19 who accessed the MIH module content, 14 participated in the on-line case discussion.

Fifteen of the 19 (79%) participants who accessed the module completed the evaluation questionnaire. These included the 14 who posted messages and one who read bulletin board messages but did not post any. List 1 summarizes their demographic and computer usage data.

Table 1 summarizes the same 15 respondents' ratings of items addressing the case discussion and the overall course. Items the learners rated most highly included relevance of the content to their practices and the prompt response of the content facilitator to their messages. One learner commented on how the content facilitator responded, "Dr. P. made sure no one felt stupid about asking a question, which is very important." They rated items addressing the bulletin board the lowest. Related comments included,

61

List 1. Respondents' Demographic and Computer-use Data, Dalhousie University, 1999

Sex	
Men	6
Women	8
Not noted	1*
Years in practice	
<5 years	4
6–10 years	1
11–20 years	6
21–30	3
Rating of computer skills	
Beginner	7
Average	6
Expert	1
What attracted you to take this course?	
Interest in headache	1
New technology	9
CME credit	0
Convenience	5
How many times did you go into the module?	
Once	2
Twice	1
3 times	1
>3 times	9
On average, how long did you spend each time you were in the module?	
<30 min	3
30–60 min	8
1–2 hours	2
>2 hours	2

* Not all of the 15 respondents to the evaluation questionnaire described in this article answered all the questions concerning demographics and computer use.

"A very frustrating experience because my computer skills were not advanced enough," and "Takes a while to get used to the bulletin board." Thirteen of the 15 respondents indicated that they wished to have more discussion-based on-line modules and would recommend this module to their peers. Supporting comments included, "The bulletin board was great once you got in" and "I think CME on-line will prove to be a Godsend for us rural physicians."

The bulletin board discussion transcript included 122 messages. The content facilit.tor posted 46 messages; the educator facilitators, 23; and the 14 learners, a total of 53. The numbers of messages posted per learner ranged from one to seven, with an average of 3.8 messages per learner. The learners posted most messages in the last two weeks of the program. Most entered the bulletin board to post messages only once, although some wrote more than one message at that time. They interacted with the case and facilitators but rarely with each other.

Analysis of the bulletin board transcript revealed four themes. These were: content (discussion of the case and questions, 80 messages), facilitative (supportive and encouraging comments, 22 messages), introductory (personal introductions, 16 messages), and administrative/technical (related to technical or logistic issues, four messages). The content expert and the learners posted all the content messages and the educators posted 16 of the 22 facilitative messages.

The content facilitator accessed the bulletin board about every second day and responded to each new learner message, giving positive feedback and stimulating critical thinking. He spent a total of about 90 minutes on-line per week. The educator facilitators accessed the bulletin board on alternate days to welcome and encourage learners and note problems. They also spent a total of about 90 minutes on-line per week. Off-line facilitator activities included contacting registrants to encourage participation, monitoring progress, and resolving problems. The content facilitator spent about 30 minutes per week in off-line activities, and the educators, about five hours per week. In addition, the facilitators undertook precourse activities to encourage participation. These included faxing participants a welcome letter, instructions, and help line information; conducting a teleconference to explain the on-line process; and posting a welcome message in the bulletin board.

Learners reported five technical problems to the help line. Four reported difficulty accessing the Web site, and one could not post a message in the bulletin board. Staff responded as promptly as possible and resolved each problem.

Discussion

Analysis of the on-line discussion confirmed that the anticipated facilitator roles were fulfilled. As content facilitator, the neurologist increased the depth and breadth of the content discussion, and the educator facilitators performed a social and supportive role by welcoming and encouraging learners. However, the neurologist, through his supportive style and prompt responses, also fulfilled a social role, and, in fact, may not have needed the assistance of the educator facilitators. A program in which the content expert is less skilled in PBL facilitation may benefit more by the addition of a skilled facilitator. Encouraging participation was an important role, expanding to off-line activities and requiring more time than anticipated. The on-line administrative/technical role was small, but it was a critical off-line function.

Despite these roles there were deficiencies in the PBL discussion. Equal learner participation is a goal of PBL,[16] but because most learners entered the discussion in the final week or two and often

Table 1. Mean Ratings of Items Addressing Case Discussion in the Bulletin Board and Overall Course by 15 Physician Participants in a CME Online PBL Program, Dalhousie University, 1999*

Item	Mean (SD)
The case content was applicable to my practice.	4.4 (0.7)
The case stimulated my thinking about patients in my practice.	3.8 (0.9)
The questions in the case clarified my understanding of the content.	3.7 (0.6)
The bulletin board was useful.	3.5 (1.2)
I received enough instruction in the use of the bulletin board.	3.6 (1.0)
I felt comfortable participating in the bulletin board.	3.6 (1.4)
Participating in discussions enhanced my understanding of the subject.	3.8 (0.9)
Discussions added value to the module.	3.8 (1.0)
The instructor responded promptly to my questions.	4.1 (0.7)
The on-line case-based format is an effective learning method for me.	3.8 (0.9)

	No. Saying Yes	*No. Saying No*
Based on this experience I would like to do more on-line modules.	13	2
Based on this experience I would recommend the module to my peers.	13	2

* Rating scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree

did not respond to messages, facilitating an ongoing discussion was difficult. Also, as few learners interacted with each other, the discussion was teacher-centered as opposed to learner-centered. A contributing factor may have been the requirement that learners post only one message in the bulletin board to receive credit. Interaction in future modules may be improved by requesting that each participant post messages weekly and respond to co-participants.

Barriers to health care professionals' adopting new communications technologies are numerous, and include the lack of adequate technical, economic, organizational and behavioral knowledge. Lessening these barriers requires intensive learning strategies.[24] Participating in an Internet discussion requires physicians to both adopt a new technology and change their learning behaviors. When asked what had attracted them to taking this course, nine of the 15 participants indicated "the opportunity to use new technology," but, with respect to their technical knowledge, seven participants rated their computer skills as "beginner." Although we provided printed instructions, a help line, and off-line support by educational facilitators, at least two registrants did not participate in this program and another two would not participate in future programs because of technology-related issues. Our findings reinforce the need for educational software that can be easily used by learners who may lack computer proficiency and have little time for or interest in mastering new technology. Providing more extensive training may increase participation, but scheduling this for busy physicians whose time is limited is difficult.

This study had several limitations. The study population was small and the learners chose to participate, so it may not represent a larger physician group. Generalizability is also limited by the lack of a control group, and although the evaluation questionnaire demonstrated face validity through a pilot-testing process, we did not test it for reliability. Of the 31 registrants, only 19 (62%) participated in the program. We learned the reasons for non-participation, important data for this study, of only four of the 12 non-participants. Ensuring reliability of the tool before repeating the study, replicating it with other populations, and being more aggressive in contacting non-participant registrants would strengthen future similar studies. In spite of limitations, this study provides insight into facilitators' roles in on-line PBL discussions and factors influencing learners' participation. It supports the view that on-line facilitators perform several roles on-line and off-line, and suggests that a challenge for facilitating PBL discussions is to promote ongoing learner–learner interaction as opposed to one-time learner–teacher interaction. Current technology hinders participation, while prompt and supportive responses by facilitators to learners' messages encourage it. All but two of the 15 learners completing the evaluation said that they would like to have more modules, indicating that the benefits outweighed the disadvantages.

Placing this small study within the context of physicians' learning, technology adoption, and behavioral change assists in considering its implications. PBL is an effective CME learning method that uses participant interaction. The Internet is a powerful tool that removes traditional barriers to both physicians' participation in CME and their interaction with co-participants, but it creates new barriers related to technology and behavioral change. We need to learn ways to overcome these barriers, a task that may become easier as communication technologies and software applications improve, and as physicians entering the workforce become more experienced in using computers.

The research reported here was funded by educational grants from the Medical Society of Nova Scotia and GlaxoWellcome Canada. The authors recognize the assistance of Dr. Jean Gray, associate dean, CME, Dalhousie University, in reviewing this article.

Correspondence: Joan Sargeant, MEd, Continuing Medical Education, Clinical Research Centre, 5849 University Avenue, Halifax, NS, Canada B3H 4H7; e-mail: ⟨joan.sargeant@dal.ca⟩.

References

1. Davis DA, Thomson MA, Oxman AS, Haynes RB. Evidence for the effectiveness of CME: a review of 50 randomized controlled trials. JAMA. 1992;268:1111–7.
2. Davis D, Thomson-O'Brien MA, Freemantle N, Wolf FM, Mazmanian P, Taylor-Vaisey A. Impact of formal continuing medical education: do conferences, workshops, rounds, and other traditional continuing education activities change physician behavior or healthcare outcomes? JAMA. 1999;282:867–74.
3. Barrows HS, Tamblyn RM. Problem-based Learning: an Approach to Medical Education. New York: Springer Publishing, 1980.
4. Doucet M, Purdy RA, Kaufman D, Langille D. Comparison of problem-based learning and lecture format in continuing medical education on headache diagnosis and management. Med Educ. 1998;32:590–6.
5. Premi J, Shannon S, Hartwick K, Lamb S, Wakefield J, Williams J. Practice-based small-group CME. Acad Med. 1994;69:800–2.
6. McCormack C, Jones D. Building a Web-Based Education System. Toronto, ON, Canada: Wiley Computer Publishing, 1995.
7. Leonardson G, Lapierre R, Hollingsworth D. Factors predictive of physician location. J Med Educ. 1985;60:37–43.
8. Dillon CL. Distance education research and continuing professional education: reframing questions for the emerging information infrastructure. J Cont Educ Health Prof. 1996;16:5–13.
9. Hersh W. "A world of knowledge at your fingertips": the promise, reality, and future directions of on-line information retrieval. Acad Med. 1999;72:240–3.
10. Barnes BE. Creating the practice-learning environment: using information technology to support a new model of continuing medical education. Acad Med. 1998; 73:278–81.
11. Bacon NC. Modernizing medical education. Hosp Med. 1999;60:54–6.
12. Komoroski EM. Use of e-mail to teach residents pediatric emergency medicine. Arch Pediatr Adolesc Med. 1998;152:1141–6.
13. Cravener PA. Faculty experiences with providing online courses: thorns among roses. Comput Nurs. 1999;17:42–7.
14. Carlton KH, Ryan ME, Siktberg LL. Designing course for the Internet: a conceptual approach. Nurse Educ. 1998;23:45–50.
15. Rosenlund CH, Damask-Bembenek B. Assessing the effectiveness of an online program. Nurse Educ. 1999;24:5–6.
16. Milstead JA, Nelson R. Preparation for an online asynchronous university doctoral course: lessons learned. Computers in Nursing. 1998;16:247–58.
17. Chan D, LeClair K, Kaczorowski J. Problem-based small-group learning via the Internet among community family physicians: a randomized controlled trial. MD Compur. 1999;16(3):54–8.
18. Barrows HS. The Tutorial Process. Springfield, IL: Southern Illinois University School of Medicine, 1988.
19. Berge ZL, Collins MP (eds). Computer-mediated Communication and the Online Classroom. Cresskill, NJ: Hampton Press, 1995:1–3.
20. Stufflebeam D. Systematic Evaluation: A Self-Instructional Guide to Theory and Practice. Boston, MA: Kluwer–Nijhoff, 1985.
21. WebCT home site: ⟨http://about.webct.com⟩. Accessed 6/30/00. WebCT Company, University of British Columbia, Vancouver, BC, Canada.
22. College of Family Physicians. Maintenance of membership site. ⟨http://www.cfpc.ca/MAINPRO/maintmem.htm⟩. Accessed 6/30/00. College of Family Physicians, Toronto, ON, Canada.
23. Bogdan RC, Biklen SK. Qualitative Research for Education. Boston, MA: Allyn and Bacon, 1992.
24. Tanriverdi H, Iacono CS. Diffusion of telemedicine: a knowledge barrier perspective. Telemedicine J. 1999;5:223–44.

63

Correlates of Physicians' Endorsement of the Legalization of Physician-assisted Suicide

KAREN D. NOVIELLI, MOHAMMADREZA HOJAT, THOMAS J. NASCA, JAMES B. ERDMANN, and J. JON VELOSKI

Although most physicians recognize a duty to provide compassionate end-of-life care, they often feel ill prepared to do so. Of particular controversy is physician-assisted suicide. Physician-assisted suicide is commonly defined as the practice of providing a competent patient with a prescription for medication for the patient to use with the primary intention of ending his or her own life. In a recent survey of approximately 2,000 U.S. physicians, 3.3% reported that they had written at least one prescription to hasten death.[1] Eleven percent reported they would write a prescription to hasten death if requested to do so under the current legal system. If legalized, 36% of the physicians would be willing to write a prescription to hasten death.[1]

Consistent with the diversity of physicians' opinions about the practice of assisted suicide, attitudes toward its legalization are also divided. When physicians in Michigan were asked to choose between legalizing or banning assisted suicide, 56% favored legalizing it, while 37% voted for a specific ban.[2]

Several studies have examined the demographic correlates of physicians' attitudes towards assisted suicide. Although age and sex were unrelated to opinions about assisted suicide,[3] race was related. Furthermore, physicians' and patients' preferences for particular approaches to end-of-life care followed similar racial patterns. White physicians were more likely than African American physicians to endorse assisted suicide in terminal care scenarios.[3] Catholic and devoutly religious physicians were also less likely than others to endorse it.[4,5]

Physicians' specialties may also help explain these differences of opinion. Oncologists were more likely to oppose assisted suicide.[6,7] Similarly, support was higher among psychiatrists than among emergency medicine physicians.[4,8,9] Only one study investigated the rationales for physicians' views on assisted suicide. One third of physicians in this study felt that it was immoral, 34% felt that it violated professional ethics, and 30% felt that it conflicted with their own religious beliefs.[10]

Since the legalization of physician-assisted suicide is an area where opinion is sharply divided, research is needed to understand the basis of physicians' beliefs about it. This study was designed to examine the extent and correlates of physicians' endorsements of the legalization of assisted suicide with regard to their specialties, sex, and opinions about certain other contemporary issues in the U.S. health care system.

Method

Graduates of Jefferson Medical College from the classes of 1987–1992 (N = 1,271) who were practicing medicine in the United States comprised the study population. Based on a search of relevant literature and two pilot studies,[11] a survey was developed that consisted of 33 items to be answered on a five-point Likert scale ("strongly agree" = 5, to "strongly disagree" = 1). The survey addressed five aspects of changes in the U.S. health care system influencing medical education, quality of care, patient referral, cost of care, ethical issues, and sociopolitical matters[11] (copies of the survey are available from the authors). The item reading "Physician-assisted suicide should be legalized" was used as the dependent variable in the present study.

The questionnaires were mailed in May 1998, followed by three reminders mailed to non-respondents at three-week intervals. Useable forms were returned by 835 physicians (66% response rate), of whom 830 responded to the item on the legalization of physician-assisted suicide. The respondents included 578 (69%) men and 257 (31%) women, with a mean age of 35.8 years. The specialties of respondents were distributed as follows: family practice, 116 (14%), general internal medicine, 85 (10%), pediatrics, 38 (5%), emergency medicine, 49 (6%), obstetrics–gynecology, 34 (4%), surgery and surgical subspecialties, 47 (6%), psychiatry, 28 (3%), hospital-based specialties (anesthesiology, pathology, and radiology), 97 (12%), medical subspecialties, 86 (10%), and other specialties and subspecialties, 255 (30%). Statistical analysis included bivariate and multivariate correlation, t test, chi-square, and z test for proportions.

Results

No significant difference was found between respondents and non-respondents with respect to gender (31% versus 33% women, respectively), age (35.8 versus 35.9 years), full-time salaried faculty appointment (14% versus 12%), and primary care practice (which was defined as family medicine, general internal medicine, and general pediatrics) (29% versus 34%).

Similarly, no difference was found for academic performance measures such as scores on the United States Medical Licensing Examinations, Steps 1–3, and clinical competence ratings provided by residency program supervisors at the end of the first postgraduate training year in three competence areas of "data-gathering and processing skills," interpersonal skills and attitudes," and "socioeconomic aspects of patient care."[1,12]

Respondents' Endorsement of Legalization of Physician-assisted Suicide. Of the 830 respondents, 284 (34%) endorsed legalization—73 (9%) "strongly agreed," and 211 (25%) "agreed"; and 340 (41%) opposed it—189 (23%) "disagreed," 151 (18%) "strongly disagreed," and 206 (25%) expressed "no opinion." The response patterns were similar for physicians who graduated in the six different cohorts.

Correlates of Endorsement of Legalization of Assisted Suicide. The endorsement rates for legalization of physician-assisted suicide were examined by the following variables:

- *Demographics.* Endorsement of legalization was unrelated to age and gender. Although the small number of African-American and Hispanic physicians in the sample was insufficient for meaningful statistical analysis. Asian physicians ($n = 48$) were significantly more likely (63%) than were whites ($n = 557$) to endorse (43%) legalization (z-test for proportions = 2.85, $p < .01$).
- *Specialty.* Orthopedic surgeons endorsed assisted suicide at the highest rate, which was 52%, followed by psychiatrists (41%), and physicians in the hospital-based specialties (40%). The lowest rates were for medical subspecialists (25%), general internists (28%), emergency medicine physicians (31%), family physicians (33%), and general pediatricians (34%). These differences in attitudes toward legalization among specialties were statistically significant ($\chi^2_{(20)} = 33.7$, $p < .05$).
- *Postgraduate ratings of clinical competence.* The physicians who endorsed legalization had been rated significantly lower by their

64

Table 1. Bivariate and Multiple Correlations and Regression Coefficients Predicting 830 Physicians' Endorsements of Physician-assisted Suicide, Jefferson Medical College*

Predictor†	Bivariate r	Regression Coefficient
Physicians should unionize to maintain the influence of their profession.	.17§	.12‡
The present paradigm of medical education does not take into account the psychosocial factors related to illness.	.15§	.15§
Government should be responsible for regulating policies that influence the quality of care.	.12§	.12‡
Learning to work in a changing health care environment should become an essential part of medical education.	.11§	.14§
Physicians involved in HMOs or other types of managed care order fewer tests than those in private practice.	.11§	.12‡
The future of health care should be based on the needs of society not on the satisfaction of physicians.	−.11§	−.09§
Physicians involved in managed care have the same dedication to their patients as physicians in fee-for-service.	−.08§	−.08‡
Intercept		2.3§
Multiple R		.30§

*Participants were 830 physicians who graduated from Jefferson Medical College between 1987 and 1992.
†Items on 33-item survey that correlated either positively or negatively with respondents' endorsement of physician-assisted suicide.
‡$p < .05$; §$p < .01$.

residency program directors in the postgraduate clinical competence areas of "interpersonal skills and attitudes" ($F_{(1,452)} = 6.25$, $p < .01$), and "socioeconomic aspects of patient care" ($F_{(1,452)} = 6.94$, $p < .01$). No significant difference was noted in the area of "data gathering and processing skills."

- *Other significant predictors of endorsement of legalization.* Bivariate correlations between responses to the item on legalization and those for other 32 items in the survey were examined. Nine items had statistically significant correlations with endorsement of legalization. A stepwise multiple regression algorithm was used, in which numerical weights assigned to responses to the item on legalization were considered as the dependent variable (criterion measure) and numerical weights assigned to the nine items of the survey that had significant correlations with responses on the physician-assisted suicide item were the independent variables (predictors). Only seven items contributed significantly ($p < .05$) to the multiple regression model, which is summarized in Table 1. Five contributed positively in that endorsement of legalization of physician-assisted suicide was associated with agreement with those items. Two contributed negatively, meaning that endorsement of legalization was associated with disagreement with those items.

As reported in Table 1, those who endorsed legalization were more likely to agree that physicians should unionize ($r = .17$, $p < .01$), that the present paradigm of medical education does not take into account the psychosocial factors related to illness ($r = .15$, $p < .01$), that government should take responsibility to regulate health care policies ($r = .12$, $p < .01$), that learning to work in a changing health care environment should become an essential part of medical education ($r = .11$, $p < .01$), and that physicians who work with managed care organizations order fewer tests than their counterparts in private practice ($r = .11$, $p < .01$).

Conversely, the physicians who endorsed legalization were more likely to disagree that the future of health care should be based on the needs of society rather than on physicians' satisfaction ($r = -.11$, $p < .01$) and that physicians in HMOs as compared with those in other settings have similar dedication to their patients ($r = -.08$, $p < .05$). The multivariate correlation was .30, $p < .01$ (see Table 1).

It is noteworthy that the responses to the item on legalization were not correlated with several other items, including the consideration of cost as an important factor in patient care decisions, physicians' support for the efforts of government to ration care, and the role that organized medicine should take with respect to social issues that can influence the well-being of society.

Discussion, Conclusions, and Implications

The findings of the present study support prior research showing that physicians hold widely disparate views regarding the legalization of physician-assisted suicide. More physicians in our study were opposed to legalization (41%) than supported it (34%), and a significant fraction of these physicians (25%) had not formed an opinion. The proportion of physicians in our study favoring legalization was similar to those in other survey work in this area.[7] Almost all respondents endorsed medical school preparation for, and subsequent provision of, compassionate care at the end of life (92%), suggesting that the differences of opinion related only to the controversial area of assisted suicide and not to caring for the dying patient in general.

Our study found that physicians in the people-oriented specialties most associated with direct and ongoing patient contact that included treatment of dying patients (general medicine, family medicine, and medical subspecialties) were less likely to endorse legalization than were technology-oriented physicians, including hospital-based specialists and orthopedic surgeons. Experience with the first year of legalized physician-assisted suicide in Oregon acknowledges the great emotional toll on physicians directly involved in its implementation.[14] The emotional burden and the acknowledged complexities in caring for dying patients may make physicians involved in this process more reluctant to endorse legalization. An interesting corollary suggested by our findings is that physicians endorsing legalization were less comfortable with their medical school training in the psychosocial aspects of care and were rated poorer in the areas of interpersonal skills and attitudes and in socioeconomic aspects of patient care in the first year of residency.

It is not known to what degree opinions about legalization are subject to modification by educational experiences during medical school. A recent study that examined medical students' views on physician-assisted suicide found that fourth-year medical students in Oregon were less likely than were fourth-year medical students in other areas of the country to be willing to provide a patient with a lethal prescription.[15] The authors suggested that a change in willingness to comply with legalized physician-assisted suicide might have occurred as a result of experience with such requests from dying patients.

Unlike many areas of medical education where knowledge is largely dependent on didactic teaching, care of the dying and attitudes towards assisted suicide are likely to be influenced primarily by personal experiences as well as the moral, ethical, and political tenets that adults bring to medical training. In addition to explor-

65

ing more closely the relationship between these personal beliefs and attitudes, an important priority for research is to determine whether attitudes towards care of the dying and physician-assisted suicide could be modified by education. Evaluation of the impact of educational experiences such as structured exposure to palliative care or rotations in a hospice service for medical students and residents would help to answer these questions. As the U.S. health care system moves from theory to practice regarding physician-assisted suicide, more research is needed to explore further the impact of legalization on physicians and their patients.

The advantages of this survey include the large sample size, gender composition, and specialty and geographic distribution of the participants that represent a broad spectrum of the population of physicians. Despite these advantages, one limitation of our study is that it ascertained physicians' views of the legalization of assisted suicide rather than their views of its practice. However, the two concepts seem logically related. The primary purpose of the survey was to gather views of multiple issues in the current health care system, including attitudes toward legalization of assisted suicide. Another limitation is that the results of this study of young physicians who graduated from a single private medical school in the Northeast may not be fully generalizable to all U.S. physicians. However, the distribution of reactions is similar to that reported in the literature.[2]

As physicians hold an influential position in the public debate on the legalization and practice of physician-assisted suicide, it is important to further understand the bases for their strong and disparate views. Futher research in this area should elucidate the political, moral, and ethical frameworks that physicians bring to this topic. Specifically, it is essential to understand the degree to which physicians' views on the legalization of physician-assisted suicide are subject to modification by medical education in general,[16] and by experiences with dying patients in particular.

Development of the foundation for this study was supported, in part, by a grant from the Bureau of Health Professions, Health Resources and Services Administration, USDHHS, under Cooperative Agreement 1 U76 MB00002-03, Centers for Medical Education Research and Policy.

Correspondence: Karen Novielli, MD, Department of Family Medicine, Jefferson Medical College, 401 Curtis, Philadelphia, PA 19107; e-mail: (karen.novielli@mail.tju.edu).

References

1. Meier DE, Emmons CA, Wallenstein S, et al. A national survey of physician-assisted suicide and euthanasia in the United States. N Engl J Med. 1998;338:1193–201.
2. Bachman JG, Alcser KH, Doukas DJ, et al. Attitudes of Michigan physicians and the public toward legalizing physician-assisted suicide and voluntary euthanasia. N Engl J Med. 1996;334:303–9.
3. Mebane EW, Oman RF, Kroonen LT, Goldstein MK. The influence of physician race, age and gender on physician attitudes toward advanced care directives and preferences for end-of-life decision-making. J Am Geriat Soc. 1999;47:579–91.
4. Schmidt TA, Zechnich AD, Tilden VP, et al. Oregon emergency physicians' experiences with, attitudes toward, and concerns about PAS. Acad Emerg Med. 1996;3:938–45.
5. Siaw LK, Tan SY. How Hawaii's doctors feel about physician-assisted suicide and euthanasia: an overview. Hawaii Med J. 1996;55:296–8.
6. Abramson N, Stokes J, Weinreb NJ, Clark WS. Euthanasia and doctor-assisted suicide: responses by oncologists and non-oncologists. South Med J. 1998;91:637–42.
7. Cohen JS, Fihn SD, Boyko, EJ, Jonsen AR, Wood RW. Attitudes toward assisted suicide and euthanasia among physicians in Washington State. N Engl J Med. 1994;331:89–94.
8. Ganzini L, Fenn DS, Lee MA, Heintz RT, Bloom JD. Attitudes of Oregon psychiatrists toward physician-assisted suicide. Am J Psychiatry. 1996;153:1469–75.
9. Roberts LW, Roberts BB, Warner TD, et al. Internal medicine, psychiatry, and emergency medicine residents' views of assisted death practices. Arch Intern Med. 1997;157:1603–9.
10. Lee MA, Nelson HD, Tilden VP, Ganzine L, Schmidt TA, Tolle SW. Legalizing assisted suicide—views of physicians in Oregon. N Engl J Med. 1996;331:310–5.
11. Hojat M, Veloski JJ, Louis DZ, et al. Perceptions of medical school seniors of the current changes in the U.S. health care system. Eval Health Prof. 1999;22:169–83.
12. Hojat M, Veloski JJ, Borenstein BD. Components of clinical competence ratings: an empirical approach. Educ Psychol Meas. 1986;46:761–9.
13. Hojat M, Borenstein BD, Veloski JJ. Cognitive and noncognitive factors in predicting the clinical performance of medical school graduates. J Med Educ. 1988; 63:323–5.
14. Chin AE, Hedberg K, Higginson GK, Fleming DW. Legalized physician-assisted suicide in Oregon—the first year's experience. N Engl J Med. 1999;340:577–83.
15. Mangus RS, Dipiero A, Hawkins CE. Medical students' attitudes toward physician-assisted suicide. JAMA. 1999;282:2080–1.
16. Batzansky B, Veloski JJ, Miller R, Jonas HS. Education in end-of-life care during medical school and residency training. Acad Med. 1999;74(10 suppl):S102–S104.

66

Learning Adolescent Psychosocial Interviewing Using Simulated Patients

KIM BLAKE, KAREN V. MANN, DAVID M. KAUFMAN, and MURRAY KAPPELMAN

The area of communication skills in adolescent medicine is emerging as a distinct and important part of the undergraduate curriculum. An appropriate level of confidence in dealing with the adolescent population is deemed a necessary educational requirement.[1] Skills in psychosocial communication with adolescents differ from those required for younger patients and adults[2–4]; they include discussing confidentiality and adolescent risk-taking activities. Simulated patients can be used effectively in teaching and evaluating of communication skills.[5–6] However, there is no report of using adolescent simulated patients to teach communication skills.

The evidence available is inconclusive regarding the teaching time required to promote retention of communication skills, although a recent review[7] suggests that one day's training or less is not effective. Long-term retention of these skills has been supported by only one paper,[8] suggesting a need to follow students over time to ascertain the effect of communication skills training.

Our study addressed two questions: (1) does feedback from a simulated adolescent patient and simulated mother lead to improvements in fourth-year medical students' psychosocial interviewing of adolescent patients? and (2) does this skill persist following the intervention?

Method

Final-year medical students (N = 68) from March 1998 through May 1999 were invited to participate, and 57 agreed. The 11 who were unavailable to participate were either interviewing for their postgraduate education, involved in presenting their own research, or unable to make the scheduled times for the simulations. Thirty five other class members were either randomly or self-selected to go to offsite locations for pediatrics, and therefore could not participate; however, this group acted as a non-randomized control arm to the study. A two-group (57 students in the intervention group and 35 in the control) prospective randomized double-bind study design was employed. The students were completing an eight-week core pediatrics rotation in a tertiary center, with seven to nine students per rotation.

Study Question 1

Intervention. Four simulated cases were developed, each comprising both a medical component (epilepsy, diabetes, attention deficit disorder, or asthma) and risk-taking activities (smoking, drugs, boyfriend issues) in which the adolescent was scripted to be involved. Nine simulated mothers and ten female adolescents (mean age 13.6 years) were recruited using established procedures.[9] Mother-and-daughter pairs were selected as this is the commonest adolescent presentation in medical practice. Young adolescents were chosen to provide a realistic presentation of this age group, which often presents a challenge to young doctors. The training for standardized feedback was achieved when all mothers reviewed a single taped scenario, scored this independently using a structured form, and then discussed the feedback they would provide the student in a group setting. The adolescents were guided by their partner mothers to give feedback, which the adolescents discussed in a focus group.

At study entry, all students signed informed consent forms. They then interviewed a simulated mother–daughter pair. The students were randomly assigned to receive immediate feedback following the pretest interview from the simulated pair (F^2), or to receive no feedback (F^1). All students conducted a second interview four weeks later using a different cas^ scenario. All students (F^1 and F^2) received feedback from the simulated pair following this post-test interview. Feedback was structured using a written modified Calgary–Cambridge guide[10] and given verbally; both interview content and process were addressed.

Measures. Three measures were taken:

1. Questionnaire. At study entry, demographic data and students' self-ratings of prior experiences with adolescent medicine, confidence in dealing with adolescent patients, and anticipated future work with adolescents were collected.
2. Pre-test. Students conducted a one-hour videotaped interview with a simulated adolescent and mother, using one of the four case scenarios, at the midpoint of their rotation. The videotaped interviews were scored by a psychologist who had been trained to reach an acceptable level of agreement with the principal investigator (KB) using the modified Calgary–Cambridge guide.[10]
3. Post-test. Four weeks later, each student conducted a second videotaped interview, using a different case scenario. Scoring was completed in the same manner as for the pre-test.

Study Question 2

Intervention. The entire final-year class participated in a mandatory ten-station OSCE prior to graduation. This was two to 12 months after participation in the study (mean 6.6 months). One pediatrics station of this OSCE tested general pediatrics knowledge (students' performances in asking about medical aspects of the case) and adolescent psychosocial interviewing (students' performances in asking about psychosocial aspects, e.g., boyfriend, alcohol, drugs). The OSCE included 35 off-site students, those not involved in the adolescent interviewing study, i.e., those who had not been videotaped and had received no feedback (F^c) and 45 of the 57 students who had completed their pediatrics rotation at the tertiary center and who had participated in the study (F^1 and F^2).

Measures. The knc vledge score and the psychosocial interviewing score on the pediatrics OSCE station were obtained from the checklists completed by the faculty examiner at the station.

Data Analysis

Study Question 1. A single psychologist, blinded to student group or time of interview, scored the tapes, using a modified Calgary–Cambridge Observation Guide.[10] The psychologist evaluated eight aspects of the encounter: how the student initiated the session, collected information, gathered information, asked the parent for time alone with the patient, dealt with the adolescent alone, and acted before and during the examination and closure. Each section yielded a global score. Within seven of the sections there were between three and ten individual items. The section used to rate when the student was alone with the adolescent included 14 psychosocial elements (i.e., boyfriend issues, smoking, and drugs).

The psychologist derived eight global ratings for each videotape. The global ratings for F^1 and F^2 students at pre- and post-test were compared using a paired t test. Regression analysis was conducted

67

using student global ratings from the eight sections of the modified Calgary–Cambridge Observaton Guide as the dependent (outcome) variable. The independent (predictor) variables were feedback, case type and simulator, gender, previous medical experience with adolescents, comfort level in relating to adolescents, future career plans, and the students' scores on the pre-test case.

Study Question 2. The knowledge score and the psychosocial interviewing score on the pediatrics OSCE station were compared among the three groups (F^0, F^1, F^2).

Results

Complete data were available for 52 of the 57 students (F^2 = 31; F^1 = 21) who completed both pre- and post-test interviews. Two tapes could not be rated, and three students did not complete the second interview.

Study Question 1. The mean pre-test scores of the group receiving feedback after their first interview (F^2) and those receiving no initial feedback (F^1) were not statistically different (72.93, SD = 9.43 versus 72.77, SD = 8.08; p = 0.95). However, the group that received feedback immediately after their first interviews (F^2) scored significantly higher on the post-test (82.81, (SD = 9.79) than did the F^1 group (76.34, SD = 9.43); p = 0.02). No significant improvement was seen from pre-test to post-test for the group receiving no initial feedback (F^1). However, the group receiving feedback (F^2) improved significantly from pre-test to post-test (p = 0.02).

Regression analysis revealed that receiving feedback was the only significant predictor (p = .021) of students' performances on the post-test case (R^2 = .10 for the complete model). The other independent variables did not significantly predict post-test performance. Analyses also were conducted to determine whether or not the particular case scenario used had a significant influence on student performance. No statistically significant influence due to case difference emerged.

Study Question 2. All students participating in the study received feedback either once (F^1) or twice (F^2). Both groups (n = 45) had significantly higher mean scores (p = .023) on the adolescent psychosocial inquiry on the final-year OSCE station (68.06, SD = 24.07) compared with the students (n = 35) who completed their core pediatrics rotation at the offsite placements (F^0) (55.71, SD = 23.16). The groups did not differ significantly (p = 0.40) in their mean scores for the general knowledge aspects of this OSCE station (F^1 and F^2) (70.71, SD = 16.88) compared with F^0 (67.53, SD = 16.69).

After the OSCE the students were asked to comment on their clerkship experience. The simulated adolescent encounters were rated as one of the most positive learning experiences in the two years of clerkship.

Discussion

The main study finding is that the important communication skill of interviewing the adolescent patient can successfully be taught to undergraduate medical students. The teaching becomes faculty-independent when the simulated patients are scripted and trained in giving structured feedback. The training period, which was a one-hour interview (experimental), followed by 20 minutes of feedback, was much less than one day, which is the time reported in the literature as necessary for effective learning of these skills. This study poses questions for further research regarding optimal training time and the best method of reinforcement. For psychosocial interviewing with sensitive questioning, clerkship seems the optimal point of instruction; however, there is little evidence to inform where training in communication with adolescents should be placed in the medical curriculum.

There are several limitations to this study. First, the sample was small, although representative of other randomized controlled trials in this field. Second, selection bias may have occurred, as the students who chose to complete their core pediatrics rotations at off-site placements were either randomly or self-selected. However, all students received the core pediatrics tutorials from the tertiary center by teleconference, along with detailed objectives. This ensured that all students received the same didactic curriculum. Third, although the study would have benefited from two independent raters, the increased cost was prohibitive. The psychologist rater was trained to use the modified Calgary–Cambridge Guide[10] and underwent a mid-study validation of his scoring. Fourth, our sample was confined to mothers and daughters; whether the results would differ with mother–son simulator pairs is unclear. Fifth, although this study provides some indication that students' psychosocial communication skills can be improved and maintained over time, follow up was less than a year. Continued tracking of these doctors would be important to see whether this mastery is maintained into the residency years. Finally, application of these results must consider resources. At our medical school, standardized patients frequently supplement current teaching activities, and are part of the diagnostic assessment of student skills throughout the medical school curriculum. Expertise to train and administer such a program is quite involved from a logistic and monetary standpoint; although available at our medical school, this may not be the case everywhere. As this educational initiative relies on a realistic portrayal and structured feedback from the adolescent, time spent in recruitment and training of the standardized patients is important.

Students overwhelmingly commented that feedback from a "real" adolescent was very helpful, as they had received little training in this area. Many of the students were very apprehensive on entry into the study, but were resoundingly positive after they had completed it.

Because of the changing nature of the hospitalized patient population, standardized patients could be used to ensure that each student has exposure to common ambulatory problems. They could help ensure uniformity in teaching and learning of basic clinical skills. Interviewing an adolescent standardized patient who is involved in risk-taking activities provides the student an opportunity to practice psychosocial interviewing in a safe setting. The immediate feedback provided by the adolescent and mother is a powerful teaching tool. The student can then return to the clinical setting to apply these newly acquired skills.

In conclusion, this randomized controlled trial has shown that final-year medical students can be taught adolescent interviewing skills and that these skills are retained for as long as a year. The teaching time required for such an intervention is short (90 minutes), and teaching can be independent of faculty once the simulators' training is completed. As the skill of talking to adolescents and their parents is an important part of physician training, we would recommend that medical schools consider this structured training for their curricula.

This research was part of Dr. Blake's AAMC Fellowship in Medical Education (1997–1998). The authors thank the members of the LRC—Nancy Ruedy, Ruth Partridge, and Susan Wakefield—for the training of simulated patients, and Marilyn Massie-Clarke for statistical analysis. Supported by a grant from the Division of Medical Education, Dalhousie University.

Correspondence: Kim Blake, MD, Dalhousie University, Department of Pediatrics, Dalhousie University 5850/5859 University Avenue, Halifax, NS, Canada B3J 3G9; e-mail: ⟨kblake@is.dal.ca⟩.

References

1. Olson, AL, Kaufman NM, Marshall SG, Woodhead J. General Pediatric Clerkship Curriculum and Resource Manual. Ambulatory Pediatric Association and the Council on Medical Student Education in Pediatrics, Chapel Hill, NC, 1995.

2. Boekeloo B, Schamus L, Cheng T, Simmens S. Young adolescents' comfort with discussion about sexual problems with their physicians. Arch Pediatr Adolesc Med. 1996;150:146–52.
3. Joffe A, Radius S, Gall M. Health counseling for adolescents: what they want, what they get, and who gives it. Pediatrics. 1998;82:481–5.
4. Wissow LS, Roter DL, Wilson ME. Pediatrician interview style and mothers' disclosure of psychological issues. Pediatrics. 1994;93:289–95.
5. Stillman PL, Regan MB, Philbin M, Haley HL. Results of a survey on the use of standardized patients to teach and evaluate clinical skills. Acad Med. 1990;65:288–92.
6. Sanson-Fisher RW, Poole AD. Simulated patients and the assessment of medical students' interpersonal skills. J Med Educ. 1980;14:249–53.
7. Aspegren K. BEME Guide No. 2: Teaching and learning communication skills in medicine—reviewing with quality of articles. Med Educ. 1999;21:563–70.
8. Maguire P, Fairbairn, Fletcher C. Consultation skills of young doctors: benefits of feedback training in interviewing as students persist. BMJ. 1986;292:1573.
9. Blake K, Greaven S. Recruiting and following adolescent standardized patients. Acad Med. 1999;74.584.
10. Kurtz SM, Silverman JD. The Calgary-Cambridge Observation Guides. Med Educ. 1996;30:83–9.

69

Have Clinical Teaching Effectiveness Ratings Changed with the Medical College of Wisconsin's Entry into the Health Care Marketplace?

DAWN BRAGG, ROBERT TREAT, and DEBORAH E. SIMPSON

Medical schools, as competitors in today's health care marketplace, have the challenge of training future physicians while increasingly relying on clinical revenues.[1] Is teaching compatible with competitive managed care in the future of health care?[2]

Skeff, Bowen, and Irby argue that teaching takes time and that its values must be re-emphasized as a core mission of medical schools.[3] Medical education researchers have reported diminishing amounts of time available for physicians' educational responsibilities to both residents[4] and medical students.[5] Student evaluations reveal that there has been less time available for them in more recent years.[6] Thus, time impacts on education have been documented, but the critical issue to be investigated is whether the quality of teaching has been compromised.

As a large, private medical school, the Medical College of Wisconsin (MCW) has not escaped the grasp of today's competitive health care environment. On December 31, 1995, the John L. Doyne Hospital (JLDH), formerly Milwaukee County General Hospital, was closed. While this facility (a primary practice and clinical teaching site) was purchased by a private adult not-for-profit hospital, it's sale nonetheless serves as a major demarcation point in MCW's transition into today's health care marketplace. Indigent care was now provided on a competitive contract basis. Our faculty formed a clinical practice group to enhance their competitive position in this evolving health care environment. Declining federal support for graduate medical education led to decreased positions in selected specialties and their associated support of medical student education. While the multi-dimensional impact of these changes on medical education, at MCW and elsewhere, will take years to analyze,[7] preliminary analysis can reveal whether the quality of clinical teaching has changed during this time period. This study, therefore, examined whether there have been changes in clinical teaching effectiveness ratings as clinicians at MCW compete for patients and revenue.

Method

The study utilized student ratings of clinical teachers from a longitudinal clinical teaching database implemented in 1992. A standard clinical teaching instrument[8] is used across participating clinical departments. The instrument contains 16 characteristics of effective clinical teaching, derived from a comprehensive review of the literature, rated using a five-point Likert scale (1 = most positive). Items address faculty interaction with students (e.g., actively involved me with patients, provided timely, constructive feedback without belittling me), ability to communicate (e.g., clear, organized, answered my questions clearly), and overall teaching effectiveness. The form is highly reliable, with a coefficient alpha of .96.

Since 1992, third-year medical students have evaluated 295 full-time clinical teachers in pediatrics, internal medicine, family medicine, anesthesiology, and general surgery. For purposes of this study, the data were divided into three time periods, using 1995 as the benchmark date for MCW's entry into health care marketplace: before-entry, 1993–94; at-entry, 1995–96; and after-entry, 1997–98 (numbers of evaluations per period = 1,327, 4,354, and 6,577 respectively).

A three-stage analytic process was used to determine whether students' ratings of clinical teaching had changed during the study period. First, the 16 clinical teaching instrument items were clustered to facilitate analysis using agglomerative hierarchical cluster analysis (HCA).[9] This method has been successfully used to cluster items on standardized tests into psychological dimensions.[10] In HCA for an n-item test, there are n solutions or clusters. In the first step, each item comprises one cluster. At subsequent steps, the procedure combines two clusters from the previous step, based upon the proximity or similarity among each possible pairing of the clusters. The smaller the proximity value, the more similar the two clusters are believed to be. The final cluster, the nth cluster, places all of the items into one cluster. By examining the two- or three-cluster solution for interpretability, a researcher can get a nonparametric perspective on groups of items that may be considered to be dimensionally distinct. Unlike factor analysis, cluster analysis is nonparametric and is a quick way to identify possible dimensions that may exist. In this study, selected clusters of clinical teaching skills were examined for internal consistency using coefficient alpha.

Using these clusters, two-way analysis of variance was performed comparing the cluster means to determine whether (1) students' ratings varied by time period; (2) students' ratings varied by item cluster; and (3) there was an interaction effect between time periods and clusters. Individual items that had been closely associated with the availability of teaching time in previous studies were then analyzed using a one-way analysis of variance to examine differences in student ratings across the three time periods.

Results

A three-cluster solution resulting from the HCA was selected for statistical and substantive reasons and to increase comparability of results with findings from prior factor-analytic studies. Ullian et al.,[11] in their synthesis of factor-analytic studies, reported that while there are varying numbers of factors, most studies suggest four solutions. The three-cluster solution was selected for this study as the two-cluster solution contained many items that did not seem to fit qualitatively and other cluster solutions contained at least one group with fewer than four items, posing a threat to internal consistency. The three clusters were examined qualitatively to assess content validity and their relationship to Ullian's four factors.

The first cluster of clinical teaching skills was labeled *supervisor/person* and contained seven items: supportive of me/had rapport with me, approachable/available, actively involved me with patients, communicated expectations, demonstrated skills/procedures to be learned, provided opportunities to practice diagnostic/assessment skills, and provided feedback without belittling me. The second cluster was labeled *physician/teacher* and contained five items: answered questions clearly, asked questions clearly, explained basis for decisions/actions, clear/organized, and clinically competent/knowledgeable. The third group, containing four items, was labeled *instructor/leader*: took advantage of teaching opportunities, enthusiastic/stimulating, responded to student-initiated learning issues, and emphasized comprehension rather than factual recall. All three item clusters, supervisor/person, physician/teacher, and instructor/leader, were found to be highly reliable (coefficient alpha = .90, .86, .80, respectively). According to Ullian et al., these three clus-

ters define the roles that clinical teachers assume in their interactions with students.

The students' ratings ranged from a minimum of 1 (most positive) to a maximum of 5 (least positive). Mean ratings across the three time periods were found to differ significantly ($p < .001$) (see Table 1). Post-hoc comparisons (i.e., Tukey test) revealed that the mean ratings for the periods were significantly different (all comparisons $p < .001$). Mean student ratings for the three clusters were also significantly different ($p < .001$). Throughout the before-entry, at-entry, and after-entry periods, physician/teacher skills were rated best by third-year students, while supervisor/person skills received the worst ratings (see Table 1). The analysis also showed an interaction between the time periods and the three groups ($p < .001$).

Mean student ratings for the three sets of skills started out positively in the first, before-entry year (see Figure 1). This was due to the fact that in 1993 faculty began to receive the first results of their clinical teaching evaluations. As reported in a prior study, when faculty receive clinical teaching evaluation results, their clinical teaching ratings improve as they immediately seek to address deficits.[12] Mean ratings for supervisor/person and instructor/leader skills increased (became worse) sharply in the second year. Mean ratings for physician/teacher continued to improve throughout the before-entry years. During the at-entry period, mean ratings for supervisor/person and instructor/leader skills continued to increase (becoming worse), but the ratings increased only gradually for physician/teacher. Supervisor/person skills peaked in 1996, the year the faculty practice plan was implemented. Mean ratings for instructor/leader and physician/teacher leveled off between 1995 and 1996. The after-entry period saw improved ratings for the three item clusters. However, none of the cluster ratings returned to the before-entry baseline level.

Of particular importance were the significant differences across time periods among the mean ratings of those characteristics associated with the availability of time. For example, mean ratings of items within the supervisor/person (e.g., supportive of me, approachable/available, actively involved me with patients) followed the increased cluster ratings. However, the ratings for "provided timely, constructive feedback without belittling me," received increasingly poor ratings across the three time periods. Analysis indicated that all four questions within this cluster were significantly different across the time periods ($p < .005$).

Table 1. Third-year Medical Students' Ratings of Physicians' Clinical Teaching Skills before, at, and after the Medical College of Wisconsin's Entry into the Health Care Marketplace, 1992–1998

	Rating*		
	Before Entry (n = 1,327)†	At Entry (n = 4,354)†	After Entry (n = 6,577)†
	Mean (SD)	Mean (SD)	Mean (SD)
Skills clusters			
Supervisor/person	1.53 (.57)	1.79 (.72)	1.76 (.70)
Physician/teacher	1.43 (.49)	1.49 (.52)	1.47 (.54)
Instructor/leader	1.48 (.54)	1.60 (.65)	1.55 (.64)
Individual items			
Supportive of me/had rapport with me	1.44 (.71)	1.69 (.90)	1.67 (.88)
Actively involved me with patients	1.37 (.67)	1.75 (.89)	1.67 (.87)
Approachable/available	1.40 (.71)	1.59 (.84)	1.57 (.86)
Provided timely constructive feedback without belittling me	1.58 (.78)	1.76 (.92)	1.80 (.93)

*Scale: 1 = most positive to 5 = least positive.
†n = number of evaluations

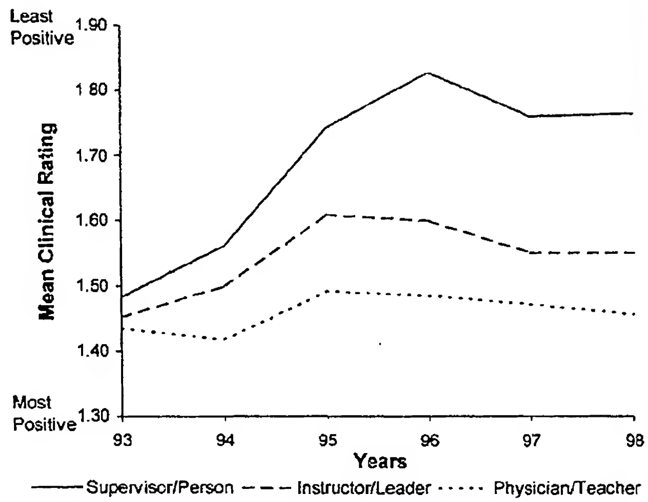

Figure 1. Students' mean ratings of physicians' clinical teaching skills across the before-entry (1993–94), at-entry (1995–96), and after-entry (1997–98) time periods.

Discussion

Longitudinal analysis of a clinical teaching evaluation data set reveals that the overall effectiveness of our clinical teaching decreased from a before-entry high at the time of entry in the health care marketplace. Over the at-entry study period, evaluations did gradually improve, but did not return to the before-entry baseline rate. However, not all item ratings were equally affected, with physician/teacher skills (e.g., clear/organized, clinically competent) showing the least change and supervisor/person skills (e.g., approachable, available, supportive of me, actively involved me with patients, provided timely, constructive feedback without belittling me) showing the largest decline. The supervisor/person skills, containing the interpersonal items, appear to have been the most profoundly affected by the entry into the health care marketplace.

Although it may be possible that students become more discriminating in their assessments of teaching and teachers over time, this study does not report ratings by the same students over time. This study used ratings by individual third-year classes for six years. In addition, student ratings were averaged over two years for each time period, thus minimizing huge class differences.

HCFA guidelines, increased pressures for clinical productivity, and accountability for cost-effective patient care have led physicians to repeatedly report that they have less time for clinical teaching. The results of this study suggest that there has also been a change in the *quality* of clinical teaching, as measured by the clinical teaching effectiveness ratings over this critical time period, a relationship requiring further study to determine causality. While it is promising that the rating results do appear to have improved following an initial decline during the at-entry period, the fact that these ratings did not return to baseline levels is distressing.

Supervisor/person skills are critical components of the teaching/learning process, as education is enhanced when there is a supportive relationship between the learner and the teacher.[11] Medical schools must prepare clinical educators with teaching skills that are effective and efficient in today's time-pressured clinical environments and implement real reward structures that recognize the value of time spent in clinical teaching if we are to maintain the quality of our clinical education.

71

Correspondence: Dawn Bragg, PhD, Assistant Director, Office of Educational Services, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226; e-mail: (dbragg@mcw.edu).

References

1. Bland CJ, Holloway RL. A crisis of mission: faculty roles and rewards in an era of health care reform. Change. 1995;30–5.
2. Farrell TA. Teaching and managed care: are they compatible in the 21st century? Arch Ophthalmol. 1997;115:251–2.
3. Skeff KM, Bowen JL, Irby DM. Protecting time for teaching in the ambulatory care setting. Acad Med. 1997;72:694–7.
4. Bolognia JL, Wintroub BU. The impact of managed care on graduate medical education and academic medical centers. Arch Dermatol. 1996;132:1078–84.
5. Nordgren R. Hantman JA. The effect of managed care on undergraduate medical education. JAMA. 1996;275:1053–8.
6. Xu G, Wolfson P, Robeson M, Rodgers JF, Veloski JJ, Brigham TP. Students' satisfaction and perceptions of attending physicians' and residents' teaching role. Am J Surg. 1998;176:46–8.
7. Xu G, Hojat M, Veloski JJ, Gonnella JS. The changing health care system: a research agenda for medical education. Eval Health Prof. 1999;22:152–68.
8. Schum TR, Yindra KJ, Koss R, Nelson DB. Students' and residents' ratings of teaching effectiveness in a department of pediatrics. Teach Learn Med. 1993;5:128–32.
9. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. New York: Wiley, 1990.
10. Stout W, Habing B, Douglas J, Kim HR, Zhang J. Conditional covariance-based multidimensionality assessment. Appl Psychol Meas. 1996;20:331–54.
11. Ullian JA, Bland CJ, Simpson DE. An alternative approach to defining the role of the clinical teacher. Acad Med. 1994;69:832–8.
12. Schum TR, Yindra KJ. Relationship between systematic feedback to faculty and ratings of clinical teaching. Acad Med. 1996;71:1100–2.
13. Palmer P. The Courage to Teach. San Francisco, CA: Jossey-Bass, 1997.

72

Six-year Documentation of the Association between Excellent Clinical Teaching and Improved Students' Examination Performances

CHARLES H. GRIFFITH III, JOHN C. GEORGESEN, and JOHN F. WILSON

With increasing fiscal pressures on academic medical centers, many institutions are moving towards mission-based financing, the notion that the clinical, research, and teaching missions must no longer depend upon cross-subsidization but must financially support themselves.[1] With this increased mission-specific accountability, there will be greater emphasis on measurable outcomes to justify the costs associated with the mission. In the realm of clinical teaching, the literature is replete with studies of qualities of excellent teachers,[2] studies of how to measure teaching,[3] and studies demonstrating that faculty development in teaching can influence clinical teachers' self-reported behaviors,[4] actual behaviors,[5] and teaching ratings.[6] However, for the most part, the fundamental outcome of teaching has been left unstudied: that is, does the quality of teaching actually influence student learning? Although this may seem a truism too obvious for investigation, despite the cherished belief of clinical teachers there is very little quantitative evidence that better teaching is associated with enhanced student learning.

We recently reported the first documentation of the association of students' learning with the relative teaching abilities of attending physicians[7,8] and residents.[9] In these studies of students and their clinical teachers over the academic years 1993–1995, we found that medical students who worked on their internal medicine or surgical clinical clerkships with our best clinical teachers scored significantly higher on post-clerkship examinations and even on the U.S. Medical Licensing Examination (USMLE) Step 2. Our findings have been replicated at the University of Michigan.[10] The only other study noting an association of teaching with learning, published in 1983, involved high school students in a remedial math class.[11] To our knowledge, this is the extent of the quantitative evidence in all the educational literature that better teaching is associated with better learning.

Our previous reports, however, had several limitations. For one, our measure of teaching "quality" was based only on students' ratings. One can argue (as we did in those articles) that the learners are the best judges of the learning climate. Even though we controlled in our analysis for prior student academic achievement (USMLE Step 1 scores), it was possible that students especially excited about internal medicine scored better on internal medicine examinations and, in their enthusiasm, rated their instructors higher, with a spurious association of examination performance and teaching rating. Second, though statistically significant, our effect sizes were modest, amounting to one-sixth to one-seventh of a standard deviation on a test, or, for example, three points on the USMLE Step 2. Third, these studies encompassed only two academic years, and a limited number of teachers and students. Because this sample was small, we included in the analysis all teachers regardless of the numbers of students they worked with, even those with few teaching ratings. Though we were gratified to demonstrate an association between teaching and learning, our results may have been attenuated by the small sample and the inclusion of in the analysis of all teachers, regardless of numbers of teaching evaluations (teachers with imprecise measures of their teaching ability).

Therefore, the purpose of this project was to refine the method of our previous studies by using a larger sample of students and attending physicians, more precise measures of teaching ability, and a way of disentangling the potential confounders of raters and teachers. Our formal hypothesis was that students who are exposed to our highest-rated attending physicians during their internal medicine clerkship will score better on end-of-clerkship examinations and on the USMLE Step 2.

Method

This work represents an extension of the data set from our previous reports,[8,9] extending the sample size from two academic years to six. The study design, a prospective cohort study, involves data on students and their attending physicians, and notes the association of the students' examination performances with the "quality" of the attending physicians to whom they were exposed. The participants were all third-year medical students at the University of Kentucky College of Medicine, over the academic years 1993–1999 and their attending physicians on the inpatient general medicine services.

To give the reader a sense of the structure of our clerkship, students in the third year spend eight weeks on general medicine inpatient services, four at our university hospital, and four at our affiliated Veteran's Affairs hospital. A team consists of an attending physician, a supervising junior or senior medicine resident, two first-year residents, and two students. Importantly, students, housestaff, and attendings are randomly and independently assigned to the services (we do not take requests by students for specific attendings). Attending physicians may be either general internists or specialists. Note that students are exposed to new and different attending physicians and housestaff in each of the two four-week components of the clerkship. Ambulatory medicine is part of a separate primary care clerkship and is not included in the study. Attending physicians usually participate in or observe daily management rounds, and have formal separate teaching rounds three times per week, ideally focused on one or two patients on the service, usually at the bedside.

Our model for how teaching might influence students' learning was not that students would be influenced by the average teaching ability of all the instructors they worked with, but rather that students' learning would be enhanced by individual outstanding instructors who, in the learning climate they engender and the inspiration they provide each day, stimulate students to be excited about clinical medicine, resulting in students' learning not only throughout the clerkship but throughout all their clinical rotations. Therefore, we explored the associations of students' learning with exposures to particularly outstanding (or poor) attending physicians, rather than with the average ability of their two attending physicians.

In our prior studies,[8,9] we simply defined "best" and "worst" attending physicians as those with the highest and lowest teaching evaluations, as rated by students. However, as mentioned in our introduction, this could lead to a confounding of teaching rating with examination performance by a student who may perform better (and rate the physician attending higher) because of interest in internal medicine. Therefore, for this study, we elected to pursue an alternative method of identifying teaching quality. We surveyed a consensus panel of third- and fourth-year residents at our institution who had also been medical students here. These individuals would have had five to six years of exposure to the clinical teachers at our university, working with a great majority of them. We also chose former students who were residents because they would be

73

TABLE 1. Least-square mean results on the NBME Subject Exam in Medicine and USMLE Step 2 for 484 Students Working with Internal Medicine Attending Physicians Rated on their Teaching as High, Neither High Nor Low, and Low, University of Kentucky College of Medicine, 1993–1999*

Attending Physicians' Ratings†	No. of Students Who Worked with an Attending of this Level	NBME Subject Exam in Medicine Score (R^2 = 0.44)		Total USMLE Step 2 Score (R^2 = 0.57)	
		Score (SD)	*p* (between Scores)	Score (SD)	*p* (between Scores)
High	219	491 (112)	.007	207 (23)	.015
Neither high not low	220	463 (112)	—	203 (22)	—
Low	45	464 (90)	—	199 (22)	—

*Least-square mean results, which represent the predicted score for a student on the test, are controlled for USMLE Step 1 score.

†High- and low-rated attending physicians were those so rated by consensus of a panel of 15 residents who had formerly been students at the University of Kentucky College of Medicine.

‡Eighteen students who had worked with both a high-rated and a low-rated attending were excluded.

most familiar with the special needs and expectations of our internal medicine clerkship. We gave these residents a list of all faculty in internal medicine who had supervised more than five medical students during the six-year period. The threshold of five students was chosen because this was the number of evaluations we calculated were needed to achieve conventional standards of reliability for our clerkship's teaching evaluation form (greater than 0.80), and it would help identify those attending physicians for whom we had precise measures of their teaching ability. We asked the residents to confidentially rate faculty "high" if they would expect them to be rated among our best teachers, "low" if they were among our worst teachers, and "medium" if they would be in between. A priori, we defined "best" attending physicians as those that were named high rated instructors by 80% of the residents (at least 12 of the 15 residents) and were not mentioned as a low-rated instructor by any resident. Conversely, realizing the tendency for learners to rate even the worst instructors at least mediocre, we defined the "worst" attending physicians as those who were rated in the low category by at least five of the 15 residents, and who were not rated high by any of the residents.

For this study, students' evaluations of attending physicians' teaching quality were also collected over the six years, as further evidence of the validity for our consensus panel opinion (one would expect the instructors who were highly rated by residents' consensus to also have high teaching ratings if the consensus process is valid). Our measure of attending physicians' teaching quality was from confidential, end-of-month student evaluations, which were completed prior to the students' receiving their grades. The form consists of 16 items on a five-point Likert-type scale (1 = strongly disagree, 5 = strongly agree). Items included ratings of teaching skills and ability, rapport with learners and patients, overall rating, and ratings of their role modeling. The coefficient alpha for the evaluation form is .96. This means that there is a high degree of internal consistency among items for rating teaching, and that the instrument is a reliable measure of teaching. However, this also means that inter-item correlations are very high, for our form .75 to .95, which is not unusual for measures of clinical teaching.[12] Because of the high inter-item correlations, we used the mean rating across all items as one measure of teacher "ability." The overall rating an instructor was assigned in our data set was the mean of all the ratings from the students he or she worked with in the six academic years.

Our analysis used multiple regression approaches from the general linear model.[13] Our dependent variables were scores on the National Board of Medical Examiner's (NBME) subject examination in medicine, taken at the end of the clerkship, and USMLE Step 2 scores. Independent variables included dummy coded variables for different categories of attending exposure (i.e., high-rated versus low-rated versus neither high- nor low-rated attending physician exposure). We also included USMLE Step 1 scores in the model as a control variable for prior student academic achievement.

Results

Data were collected from 502 third-year medical students (100% of students) over the six academic years. We excluded 18 students who had worked with both a high-rated and a low-rated instructor (as our model was less clear about how this interaction might influence student learning), for a final sample of 484. A total of 46 attending physicians had more than five student evaluations over the six-year period and were included in the list that the consensus panel rated.

Overall, ten faculty met the criteria to be rated "high." Eight of the ten were rated high by all residents, and the other two by 13 and 14 residents, respectively. Four of the ten were general internists. Eight were men and two were women, which reflects our faculty demographics. Five faculty met our consensus criteria for a "low" rating, including one general internist and one woman.

Teaching evaluations were received from 96% of the students. The overall mean teaching rating of the teachers rated high was 4.68 on the five-point scale (SD = 0.22, range 4.23–4.94). For the "medium" group, the mean teaching rating was 4.34 (SD = 0.32, range 3.4–4.92). For the "worst"-rated attending physicians, the mean rating was 3.56 (SD = 0.48, range 3.06–4.21). Mean differences between groups were highly statistically significant ($p <$ 0.001). Forty-five students had had exposures to at least one low-rated and no high-rated attending physician; 219 had had exposures to at least one high-rated and no low-rated attending physician; and 220 had had exposures to neither a high- nor a low-rated attending physician. Our high-rated attending physicians were more often attending physicians on the general medicine inpatient services than were the low- or medium-rated faculty, hence the disparity in numbers of students per faculty.

Table 1 presents the least-square mean scores on the post-clerkship NBME subject examination in medicine and on USMLE Step 2, depending on exposure to high-, low-, or medium-rated instructors (least-square means are predicted means adjusted for USMLE Step 1 scores). As can be seen, students who worked with at least one of our consensus panel's highly rated instructors scored significantly higher on the post-clerkship NBME examination in medicine and USMLE Step 2.

Conclusions

Our findings once again confirm the association of better clinical teaching with better student examination performance, demonstrating in a quantitative fashion the outcomes of teaching. The effect sizes in this current project are much more substantial than those in our prior reports, amounting to one-fourth to one-third of a standard deviation, or, for example, up to seven or eight points on USMLE Step 2, versus the one-sixth to one-seventh standard deviation effect sizes of our prior reports. We attribute our stronger conclusions to the more refined method of this current project.

First, our previous reports included all instructors, regardless of the numbers of students they had taught, and therefore all faculty were eligible to be included in the high- or low-rated category even if they had few student evaluations. For example, we may have included in our high category those faculty with only two or three ratings that were all high, when over time their ratings might have regressed towards a more stable mean that did not qualify them as such. In essence, we were categorizing some instructors as better or worse by using imprecise measures of their teaching ability. This imprecision would tend to add background "noise" to the analysis, attenuating our findings and effect sizes. Second, we disentangled learner outcomes from ratings by learners with our residents' consensus panel. As shown, attending physicians who were rated highly by the residents' consensus panel had significantly higher teaching ratings than did the medium- and low-rated instructors. Our previous method, relying on categorization solely by teaching rating, may have led to the exclusion of some otherwise excellent clinical teachers simply because they did not quite meet the "top 20% of student evaluations" we had required in our previous report to be considered a highly rated instructor.

Our findings seem to indicate that clinical teaching has an influence on outcomes, such as performance on USMLE Step 2. One might wonder how a short four-week exposure in a single discipline could influence USMLE Step 2 scores to such a degree, given that USMLE Step 2 comprises a wide variety of disciplines. Our answer is suggested by our model. From our experience as learners, the influence of a single outstanding instructor on one's approach to learning should not be underestimated. We suspect that the best teachers do not necessarily impart more factual information (facts which may be obsolete in a few years), but rather they engender a learning climate that makes learning fun, enjoyable, and exciting. They may do this by their example, by modeling the process of lifelong learning, by the joy they bring to their teaching, or by combinations of qualities such as these. Regardless, the learner's approach to learning is in some fundamental way changed, carrying over to the other clerkships and, we hope, to residency and beyond. Further studies should investigate the influence of outstanding teachers on life-long learning.

Several limitations to our study should be kept in mind as one interprets our results. This is a single-institution, single-discipline study, and certainly national studies are needed to assert the generalizability of our findings, as well as studies in other disciplines. In addition, our study focused on but one outcome measure, students' performances on NBME-type examinations, which measure but one aspect of clinical ability (knowledge). Future research should investigate the influence of teaching on other student outcomes, such as clinical skills, attitudes towards patients and the professior., and doctor–patient communication and relationships. Finally, this project's method did not lend itself as well to measuring the influence of residents' teaching on students' outcomes, so further studies are needed.

Nevertheless, despite these limitations, we conclude that attending physicians' teaching quality can have a measurable impact on students' examination performances. We therefore believe it is possible to begin considering learners' outcomes as an important measure of faculty's teaching ability, perhaps (with more study) an important addition to teaching portfolios and promotion dossiers. But even more, we believe our findings add to the growing literature on the critical importance of the educational mission that indicates students' learning would be jeopardized if the educational mission were to be compromised for fiscal reasons.

This research was supported by a grant from the National Board of Medical Examiners Research Fund, 56-9798.

Correspondence: Charles H. Griffith III, MD, MSPH, Department of Internal Medicine, University of Kentucky College of Medicine, Kentucky Clinic K512, Lexington, KY 40536-0284.

References

1. Watson RT, Romrell LJ. Mission-based budgeting: removing a graveyard. Acad Med. 1999:74:627–40.
2. Irby DM. What clinical teachers in medicine need to know. Acad Med. 1994;69: 333–42.
3. Speer AJ, Elnicki DM. Assessing the quality of teaching. Am J Med. 1999;106: 381–4.
4. Skeff KM, Stratos GA, Bergen MR, Sampson K, Deutsch SL. Regional teaching improvement programs for community-based teachers. Am J Med. 1999;106:76–80.
5. Wilkinson L, Sarkan RT. Arrows in the quiver: evaluation of a workshop in ambulatory teaching. Acad Med. 1998:73(10 suppl):S67–S69.
6. Skeff KM, Stratos GA, Berman J, Bergen MR. Improved clinical teaching: evaluation of a national dissemination program. Arch Intern Med. 1992;152:1156–61.
7. Blue AV, Griffith CH, Wilson JF, Sloan DA, Schwartz RW. Surgical teaching quality makes a difference. Am J Surg. 1999;177:86–9.
8. Griffith CH, Wilson JF, Haist SA, Ramsbottom-Lucier M. Relationships of how well attending physicians teach to their students performance and residency choice. Acad Med. 1997;72(10 suppl):S118–S126.
9. Griffith CH, Wilson JF, Haist SA, Ramsbottom M. Do students who work with better housestaff in their medicine clerkship learn more? Acad Med. 1998;73(10 suppl):S57-S59.
10. Stern DT, Gill A, Gruppen LD, Grum CM, Woolliscroft JO. Do the students of good teachers learn more? Abstract presented at Research in Medical Education Annual Meeting, Washington, DC, November, 1997.
11. Friedman M, Stamper C. The effectiveness of a faculty development program: a process-product experimental study. Rev Higher Educ. 1983;7:49–65.
12. Marriott DJ, Litzelman DK. Student's global assessments of clinical teachers: a reliable and valid measure of teaching effectiveness. Acad Med. 1998;73:572–4.
13. Cohen J, Cohen P. Applied multiple regression correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

75

When Residents Talk and Teachers Listen: A Communication Analysis

JUDY L. PAUKERT

Communication is not only an exchange of ideas but also a form of social behavior that negotiates relationships. How two parties talk with each other reveals their relative status, level of rapport, and value for each other. Not surprisingly, the power that a speaker derives from his or her status may jeopardize a conversation. The teacher's role, particularly as evaluator, often leads the teacher to dominate conversation with the learner. In one-on-one teaching and other dyadic interactions, the less powerful party expects to adapt to the dominant party's speech and initiations.[1] When adaptation is extreme, communication is authoritarian; the more powerful party rejects the less powerful party's speech by interrupting, taking over, or monopolizing the conversation. When adaptation is minimal, communication is autonomous; the more powerful party encourages the other to dominate or lead the conversation by verbal and nonverbal behaviors.

Although several studies have demonstrated that residents perceive autonomy as important to their learning,[2,3] research about the effects of interactions between residents and attending physicians on the development of clinical independence has produced contradictory findings.[4,5] No study has described how autonomy emerges from communication between residents and their teachers. Analysis of conversations between teachers and learners has generally been limited to determining the amount and duration of contact. Most content analysis has focused on the topics discussed and categorization of utterances.[6] However, analysis of another dyadic interaction, physician–patient communication, has identified several distinct patterns based on communication control and verbal dominance.[7,8]

An in-depth examination of communication patterns between the physician–teacher and the physician-in-training may increase our understanding of the types of interactions that help residents learn. This study analyzed how preceptors and residents interact during teaching encounters in ambulatory pediatrics primary care settings. This study focused primarily on autonomous communication when residents dominate the conversation.

Method

The Institutional Review Board approved this study, which was conducted in the continuity care clinics of the general pediatrics residency program at Baylor College of Medicine, Houston, Texas. The study involved both academic and community (private practice) sites. Preceptors were selected based on diversity of teaching reputation, teaching and pediatrics experience, interpersonal skills, and practice setting (solo to large group). The final sample was made up of six academic and seven community preceptors. Four to nine clinical teaching encounters were observed and audiotaped for each preceptor. Each encounter was a unique opportunity to capture a communication pattern. In all, 76 preceptor–resident interactions were analyzed using the grounded-theory method.[9] An experienced educator re-examined and independently coded a portion of the transcribed encounters. Intercoder agreement was about 95%. Participating preceptors were also asked to confirm the analyses.

Results

The encounters included acute care, follow-up, and well-child visits and involved first-, second-, and third-year residents. Four distinct patterns of communication were identified based on conversational input and verbal dominance. Of 76 interactions, 54 (71%) showed a conversational balance between speakers: 47 mutual (high preceptor and high resident input) and seven default (low preceptor and low resident input). The remaining 22 interactions showed imbalances between speakers: 15 autonomous (high resident and low preceptor input) and seven authoritarian (high preceptor and low resident input).

Almost 20% of the interactions were classified as autonomous. Of these, 12 (80%) occurred in community settings. Two academic and four community preceptors engaged in autonomous interactions. No preceptor relied on autonomous interactions exclusively, although one academic and one community preceptor used only authoritarian communication. Thus, 11 of the 13 preceptors used more than one communication pattern.

Further analysis of autonomous interactions revealed specific preceptor behaviors. In every autonomous interaction observed, the preceptor recognized the resident's "expertise" and allowed the resident to dominate communication during the interaction or, at least, the conversation about the patient. Generally, the preceptor's approval resulted from the preceptor's identifying the resident's level of understanding as appropriate for a case. The examples reported in the following sections represent behaviors observed across the series of autonomous interactions.

Probing Questions. Preceptors used probing questions to assess residents' understanding. In one community encounter, a first-year resident presented an 8-year-old child complaining of nighttime coughing and congestion related to physical exertion. After listening to a concise but detailed exposition of subjective and objective findings, the preceptor asked, "What do you think of his sequelae to his respiratory infection, exercise cough, and that kind of thing?" The residents' response confirmed a level of understanding appropriate for diagnosing the patient's condition:

> R: Well, it's pretty likely that he has some kind of twitchy airway. He's had a recent infection and recent irritation to his lung and he just had another little cold. So anything that he might get on top of it might cause him to have a little bit tighter air flow. So maybe at night, [and] that might be one aspect of reactive airway disease, especially when he's active.

In another encounter, an academic preceptor used variations of the sample probe throughout a third-year resident's presentation of a 5½-year-old girl with Angelman's syndrome and a febrile seizure disorder. The preceptor began probing after the resident had finished presenting the subjective findings:

> R: Also [she was] seen by Neurology for a history of febrile seizures seen with infections. She's been on Depakene. She has been on several medicines. First Dilantin liquid, then Dilantin tablets that were crushable, but she had a lot of drooling and would drool out most of it. Depakene, first the sprinkles and now the elixir.

> P: They really thought this was a seizure disorder and not just febrile seizures?

> R: Thought so. [Looking through chart for Neurology entry] Impression is seizure disorder, febrile. Recommended an EEG [electroencephalogram] which Mom said was done but was not a good study, and hasn't been seen in the [Neurology] Clinic in about two years.

76

> Mom wants to take her off the Depakene and I told her that I would want her to be seen by the physicians [neurologists] here.

In this encounter, the resident responded to the question by giving an "expert" answer: citing the chart entry made by Neurology. Then, the resident elaborated by expressing the need and rationale for obtaining a blood test to determine whether (1) the current dosage of anticonvulsant was therapeutic and (2) stopping the medication would do no harm. The preceptor's remarks to the resident's concise assessment and plan for this complicated patient with multiple medical problems signaled support of the resident's autonomy:

> R: So, impression is history of Angelman's syndrome, developmental delay, history of complex febrile seizures, left exotropia, and [patient] would probably benefit from visits back to her multiple subspecialists.

> P: So, we're going to check her Depakote level today and then maybe decide [about taking her off Depakote]. And you want her to go to Neurology as well.

The preceptor signified concurrence with the resident's plan by using "we" to show agreement ("we're going to check her Depakote level today and then maybe decide"). Similarly, the preceptor confirmed the resident's autonomy and dominance by using the pronoun "you" ("you want her to go to Neurology as well").

Inference. In other encounters, the preceptor inferred the resident's level of knowledge and understanding from the organization, thoroughness, and conciseness of the case presentation, a finding also demonstrated by Irby.[12] The case presentations of the only third-year resident observed in the community setting (four interactions) were so well organized and articulated that the preceptor rarely commented other than to agree with the resident's findings. The resident almost monopolized the conversation, with a smooth, confident, and complete case presentation in SOAP format (subjective–objective–assessment–plan). The preceptor spoke only when the resident paused and showed agreement by using minimal reinforcers, such as "all right," "okay," and "that sounds right." Conversationally, minimal reinforcers cue the speaker that the listener is involved and following the speaker's thoughts.[11] In a clinical interaction, these utterances also cued the resident that the preceptor was willing to allow the resident to dominate the talk and the encounter.

Besides smoothness, proper terminology, and adherence to SOAP format, other characteristics of the presentation permitted a different community preceptor to assess a first-year resident's level of knowledge. In this encounter, the resident was confident but more relaxed in style and language than the previously described third-year resident. Satisfied with the resident's presentation of the subjective and objective findings, the preceptor asked for the resident's assessment and plan. The resident replied, "Her right TM looks like really white. I just, I guess that there's pus behind it. . . . It looks way different than the left side. . . . So right otitis. And since she's never had any problems before, just do amoxicillin." Although not eloquent, the resident's response brought agreement from the preceptor and closed the encounter.

Admittedly, without probing the extent of a resident's knowledge, a preceptor might wrongly infer a resident's understanding was appropriate. By engaging in the behavior of "showing," that is, confidently presenting findings and knowledge of certain entities of a case, a resident might hide actual deficiencies of other entities within the same case. The preceptor's own knowledge of a resident's past performance, particularly for a disease or family of diseases, may prevent some mistakes. For example, in one encounter, an academic preceptor asked a second-year resident to limit the case presentation and "give the big points." The resident condensed the subjective and objective findings into two sentences: "These kids are here for well-child checks. The bottom line is that the older one has lice and ringworm, and the younger one has only lice." The resident's response probably tapped into two important pieces of information available to the preceptor. First, the preceptor knew what a second-year resident should know about lice and ringworm, both common pediatric problems. Second, the preceptor knew this resident specifically from interactions over the preceding two years. Despite this knowledge, the preceptor listened attentively throughout the resident's speech and offered minimal reinforcers like "oh no" and "okay."

Nonverbal Behaviors. In the 15 autonomous interactions, all preceptors listened attentively and used verbal and nonverbal behaviors to indicate that they followed the residents' reasoning and talk. Nonverbal behaviors, such as eye contact, facial expressions, head nods, and alert body posture, more accurately disclose how well a party is listening to a conversation than do verbal behaviors.[12] Another important nonverbal behavior observed was the control of the patient's chart. In most autonomous interactions, the resident controlled the patient's chart. Controlling the chart prevented the preceptor from reading the chart during case presentation and diverting the conversation to an unrelated chart entry, behaviors observed in the seven authoritarian interactions characterized by preceptor dominance. In both autonomous and authoritarian interactions, the dominant speaker controlled the patient's chart.

Absence of Teaching Scripts. Preceptors did not use a clinical teaching script[10] in any autonomous interaction. Fatigue or time of an encounter, such as the last encounter of a late-running clinic, might have affected a preceptor's decision not to use a clinical teaching script, but instead to permit the resident's autonomy. Arguably, a preceptor who failed to assess a resident's understanding might have missed an opportunity to use a teaching script.

At least one autonomous interaction exemplified how a clinical teaching script might have been less effective than the teaching created by follow-up of the case itself. This encounter was the follow-up visit of a "fussy" 2-week-old child, who had been consoled only by feeding when initially seen by the same academic preceptor and first-year resident. The initial teaching encounter, which corresponded to the first visit, contained a clinical teaching script regarding diagnosis and management of suboptimal weight gain in a baby. At that time, the preceptor and the resident negotiated a treatment plan to increase the number of feedings and rule out gastroesophageal reflux as an organic cause for fussiness and poor weight gain. Between the initial and follow-up visits, the patient had been x-rayed and started on appropriate medication for reflux by another preceptor and resident. By the follow-up visit (and second teaching encounter), the resident was able to learn from the case itself and to see the results of the patient's treatment. The resident's speech reflected understanding of the case and delight with the patient's improvement:

> R: She did not cry once, the whole time I was in there.

> P: You're kidding. Was this the same baby we saw one week ago [laughing]?

> R: No. I was thinking, this was a completely different baby. You know, when I examined her, I did everything. You know? She's alert, looking around. I mean she wasn't lethargic. She was fine, screaming at the top of her lungs.

> P: Wonderful!

> R: Mom says that she's not fussy at home the way she had been. This is the way she is like at home as well. [She] is breastfeeding about every 2 to 2½ hours, 15 minutes on each breast. Seems satisfied after each feed. The feedings are going better since she started the Zantac and Cisapride.

Observing the post-treatment changes in the patient reinforced the preceptor's earlier teaching script. This pair of encounters also demonstrated the benefits of the continuity care experience in

77

which a resident develops and maintains a continuing physician—patient relationship with a panel of patients. The preceptor's experience with a resident and that resident's panel of patients may increase the likelihood that the preceptor will allow the resident greater autonomy in the teaching interaction.

Conclusions

Autonomous preceptor–resident communication is characterized by high resident and low preceptor input and preceptors' behaviors that confirm and recognize the residents' speech. The preceptor assesses the resident's level of understanding in a particular case through questioning or inference from the organization, thoroughness, conciseness, and confidence of the resident's speech. Twelve of the 15 autonomous interactions occurred in community settings. The reasons for this difference in the rate of autonomous interactions between academic and community settings were not identified. Conversation is a response not only to a person but also to environmental conditions, such as the available time and space for teaching. Economic factors may encourage community preceptors to permit autonomy rather than provide more directed teaching.

Academic and community preceptors were alike in other respects, particularly the use of multiple communication patterns. This finding suggests a spectrum of the relationships between teacher and learner that encourage different conversational behaviors. The use of multiple communication patterns may indicate "scaffolding," an overarching process within the preceptor–resident relationship in the continuity care experience.[13]

Scaffolding refers to techniques that support learners in their efforts to solve difficult problems or perform difficult tasks. For a novice resident, a preceptor may behave authoritatively to provide maximum support by modeling desired behaviors, such as how to perform an examination or give anticipatory guidance. As the resident's experience and skill in relating to the preceptor and patients increase, the need for support decreases. Thus, the preceptor behaves less authoritatively, and more collaborative and autonomous interactions occur.

Scaffolding requires the preceptor to know what support a resident truly requires. The space and time available for teaching may increase the preceptor's reliance on autonomous interactions, even when there is a recognizable teaching moment. In this study, two extremes were found: some first-year residents were involved in autonomous interactions and some third-year residents in authoritarian interactions. Possibly, preceptors select the amount of support to give a resident based on the resident's specific experience with a problem. Whitman and Schwenk[14] seem to advocate "selective" scaffolding by suggesting that medical teachers alternate between assuming active and passive roles, depending on learners' needs. Training preceptors in scaffolding techniques is not likely to eliminate default communication patterns, particularly when fatigue undermines conversation.

This study is limited because it was performed in a pediatrics setting. Despite its setting, the study has potential implications for all clinical teaching that involves one-on-one interactions between learner and teacher. It is probable that the communication patterns identified may be observed in other practice areas because this study did not limit participation to exemplary teachers and sampled for diversity. The effect of observation on participants' behaviors cannot be entirely discounted. However, over a clinic session, both preceptors and residents seemed to forget that they were being observed.

Future studies should delineate how different communication patterns affect teaching and learning. The relationship of the teaching interaction to future physician—patient communication also deserves investigation. For example, do the ways that preceptors talk with residents predict the ways that residents talk with patients?

Teaching preceptors active listening and supportive verbal and nonverbal behaviors may benefit both patients and residents. When preceptors carefully attend to their residents' presentations, by listening and acknowledging the residents' concerns and opinions, they model how physicians should attend to their patients' conversations, by listening and acknowledging patients' concerns. Studying the conversation of clinical teaching should illuminate how physicians-in-training learn the conversation of healing.

Correspondence: Dr. Paukert, Mail Code 7737, Department of Surgery, The University of Texas Health Science Center, 7703 Floyd Curl Drive, San Antonio, TX 78229-3900. Reprints are not available.

References

1. Cissna KNL, Sieburg E. Patterns of interactional confirmation and disconfirmation. In: Wilder-Mott C, Weakland J (eds). Rigor and Imagination. Westport, CT: Praeger, 1981:230–9.
2. Ullian JA, Bland CJ, Simpson DE. An alternative approach to defining the role of the clinical teacher. Acad Med. 1994;69:832–8.
3. Stritter FT, Baker RM. Resident preferences for the clinical teaching of ambulatory care. J Med Educ. 1982;57:33–41.
4. Williamson HA, Glenn JK, Spencer DC, Reid JC. The development of clinical independence: resident–attending physician interactions in an ambulatory setting. J Fam Pract. 1988;26:60–4.
5. Knudson MP, Lawler FH, Zweig SC, Moreno CA, Hosokawa MC, Blake RL. Analysis of resident and attending physician interactions in family medicine. J Fam Pract. 1989;28:705–9.
6. Irby DM. Teaching and learning in ambulatory care settings: a thematic review of the literature. Acad Med. 1995;70:898–931.
7. Roter DL, Stewart M, Putnam SM, et al. Communication patterns of primary care physicians. JAMA. 1997 . 7:350–6.
8. Emanuel EJ, Emanuel LL. Four models of the physician-patient relationship. JAMA. 1992;267:2221–6.
9. Glaser B, Strauss A. The Discovery of Grounded Theory: Strategies for Qualitative Research. Chicago, IL: Aldine, 1967.
10. Irby DM. How attending physicians make instructional decisions when conducting teaching rounds. Acad Med. 1992;67:630–8.
11. Brownell J. Building Active Listening. Englewood Cliffs, NJ: Prentice Hall, 1986.
12. Grove TG. Dyadic Interactions. Dubuque, IA: William C Brown Communications, 1991.
13. Ormrod JE. Human Learning. Englewood Cliffs, NJ: Prentice-Hall, 1995.
14. Whitman NA, Schwenk TL. The Physician as Teacher. 2nd ed. Salt Lake City, UT: Whitman Associates, 1997.

78

The Relationship between the Nature of Practice and Performance on a Cognitive Examination

JOHN J. NORCINI and REBECCA S. LIPNER

Certification by a specialty board affiliated with the American Board of Medical Specialties is one of the most widely used markers of physician competence in the United States.[1] To enhance the meaning of this credential and in recognition of the need for periodic reassessments of physicians, the specialty boards have time-limited their certificates. To maintain certification, most of the boards have developed programs that incorporate (1) a check of credentials, (2) self-evaluation and/or continuing medical education, and (3) a secure written or computer-based examination. The secure examination is considered an integral component of certification because it provides assurance that physicians are keeping up with changes in medical knowledge and that they possess the ability to successfully manage patients' problems that are important but rarely encountered in practice. Moreover, from patients' and payers' perspectives, a secure examination lends more credibility to the certification process.

Despite these benefits, the secure examination is contentious because many believe that it tests only the ability to recall factual knowledge and, as such, bears little relationship to the day-to-day practice of medicine.[2] However, the growth of electronic records and databases has made it possible to begin to address this concern by comparing certification status and test performance with aspects of practice such as volume, process of care, and patients' outcomes.

There is considerable evidence that physicians who treat large numbers of patients with a particular condition generally provide better care for such patients. Volume is directly associated with patients' outcomes, regardless of the discipline or procedure.[3] Therefore, the relationship of practice volume to examination scores is an important part of any test-validation effort. A study involving the first geriatric medicine certifying examination indicated that the number of geriatric patients seen in practice was positively correlated with examination scores.[4] Likewise, a study involving cardiologists indicated that performance on cardiovascular graphics questions was positively correlated with experience.[5] Specifically, scores on the interpretation of echocardiograms were correlated with the numbers of echocardiograms interpreted in practice or training. Similarly, scores on the interpretation of arteriograms and ventriculograms were correlated with the numbers of angioplasties performed. More recently, a study of a cognitive recertification examination in critical care medicine showed that scores were related to the amounts of time physicians spent in the direct care of critically ill patients. This relationship persisted even after statistically removing performance on the initial certifying examination in the same discipline.[6]

There is also evidence that certification status and examination performance are related to the process of care. A study of physicians in Quebec compared consultation rates, inappropriate prescribing for the elderly, and mammography screening rates with licensing examination scores based partly on cognitive tests.[7] Physicians with higher scores referred more patients for consultation, prescribed fewer inappropriate drugs and more disease-specific medications for symptom relief, and appropriately referred more women for mammography. Similarly, a study of certified and non-certified internists found differences in preventive care services favoring the certified physicians.[8]

Although it is difficult to measure good practice outcomes well, some progress is being made in using available data in the validation of cognitive examinations. A recent study investigated whether there were differences among certified and self-designated cardiologists, internists, and family practitioners in the mortality of their patients with acute myocardial infarction.[9] Data for all myocardial infarctions for calendar year 1993 in Pennsylvania were analyzed. Certification was associated with a 15% reduction in mortality irrespective of specialty, and after taking account of severity of illness, hospital characteristics, patient volume, and years since graduation. Similarly, Ramsey and colleagues found differences in some outcomes that favored the certified physicians,[5] and there are a few studies with similar positive results in other specialties.[10-12]

Given the significance of the topic and the need for additional investigation, the purpose of this study was to extend previous work by exploring in more detail the relationship between examination performance and the nature of practice. Specifically, candidates for recertification in critical care medicine supplied information, via a practice survey, about the amounts of time that they spent in the care of patients with cardiovascular and pulmonary problems (i.e., practice volumes). Moreover, they rated the complexity of the problems they saw. These practice data were compared with performances on the items from the examination that dealt specifically with cardiovascular and pulmonary problems to determine whether patient complexity, in addition to patient volume, was associated with test scores.

Method

Participants. The data are based on the candidates who attempted the 1997 and 1999 recertification examinations in critical care medicine and responded without error to the practice survey. All of these candidates had time-limited initial critical care certificates. Ninety-nine percent and 93% of their certificates expired in 1997 and 1999, respectively. In 1997, the average examinee had been certified in internal medicine in 1979 (18 years, SD = 4 years), and these physicians spent most of their time in direct patient care (mean = 70%, SD = 26%). In 1999, the examinee group's average candidate had been certified in internal medicine in 1982 (17 years, SD = 4 years), and these physicians also spent the majority of their time in direct patient care (mean = 72%, SD = 26%).

Examinations. The 1997 and 1999 critical care medicine recertifying examinations each consisted of 120 single-best-answer questions, all of which were asked in the context of a clinical problem and most of which required synthesis and judgment to reach the correct response. Consistent with the purpose of the tests, the content focused on well-established principles of patient care that should be known without consulting medical resources.

The questions were written by a test committee of experts, but before these questions were selected for the examinations, they were sent to critical care practitioners who rated them for relevance to practice. The examinations had average relevance ratings of more than 4 on a five-point rating scale, where 5 denoted "very relevant." The same items appeared on the 1997 and 1999 critical care medicine *initial* certifying examinations.

This study concentrated on the 1997 and 1999 exams' cardiovascular and pulmonary disease questions because problems in these areas were frequently encountered in practice by candidates, and they were the largest subsets of items on the examination. In ad-

79

dition, these examination years had substantial numbers of candidates taking the examination. Table 1 presents the numbers of items and their means (SD) for the 1997 and 1999 critical care medicine recertification examinations. For this study, subtest scores were reported on the raw score scale.

Since scores on initial certifying examinations are related to features of residency and fellowship training as well as fund of medical knowledge, the *initial* certifying examination in critical care medicine was used as a statistical control in this study.[13] The analyses took account of the scores on this examination and attributed effects to other variables if they made independent contributions to the explanation of the performance. The scores on the initial certifying examination had been standardized against a national group with a mean of 500 (SD = 100) and were equated over years using a common-item linear equating technique. Table 1 presents the means and standard deviations for each cohort.

Survey. When physicians applied for the examination, they were asked to supply information about their practices. Specifically, for each of the specialties of medicine, they were asked what percentage of time they spent with patients. Physicians whose responses did not amount to 100% were removed from the analysis. Candidates were also asked to rate on a five-point scale (where 1 was "not very complex" and 5 was "very complex") the complexity of the cardiovascular and pulmonary disease cases they managed. Finally, to test the joint effect of time and complexity, the ratings were multiplied by percentage of time spent in an area. Table 1 presents descriptive data for these variables.

Procedure. The data were submitted to four separate stepwise linear regressions, two (1997 and 1999) for cardiovascular disease and two (1997 and 1999) for pulmonary disease. The dependent measures were the cardiovascular and pulmonary subscores on the critical care medicine recertifying examination and the independent variables were (1) score on the initial critical care medicine certifying examination, (2) the frequency of patients' problems encountered in the area, (3) the complexity of those problems, and (4) the interaction of the two factors (i.e., frequency times complexity).

Table 1. Descriptive Data and Regression Result for Physicians Who Took the 1997 and 1999 Critical Care Medicine Recertifying Examinations and Completed a Survey on Their Practices' Characteristics

	1997 (n = 510)	1999 (n = 334)
Certifying exam scores, mean (SD)	572 (59)	568 (58)
Cardiovascular disease questions	(n = 38)	(n = 24)
Number right, mean (SD)	30.8 (3.3)	19.5 (2.3)
Complexity rating, mean (SD)	3.1 (1.3)	3.3 (1.3)
% of time, mean (SD)	10% (12)	11% (13)
Interaction of frequency and complexity, mean (SD)	37.7 (52.5)	43.8 (59.5)
β coefficients		
Constant	16.50	10.66
Initial certifying exam subscore	.025	.015
% of time	NS	NS
Complexity of problems	NS	NS
Time–complexity interaction	.007	.008
Pulmonary disease questions	(n = 14)	(n = 26)
Number right, mean (SD)	11.5 (1.6)	20.0 (3.2)
Complexity rating, mean (SD)	4.5 (1.0)	4.3 (1.1)
% of time, mean (SD)	35% (24)	33% (24)
Interaction of frequency and complexity, mean (SD)	158.4 (113.4)	150.7 (112.3)
β Coefficients		
Constants	5.92	1.83
Initial certifying exam subscore	.009	.026
% of time	.007	.017
Complexity of problems	NS	.71
Time–complexity interaction	NS	NS

Results

In predicting the cardiovascular disorder subscore on the 1997 recertification examination, the critical care medicine certifying examination entered first into the regression equation (R^2 change = .18, t = 10.44, p < .001), followed by the interaction of the frequency and complexity of patients with cardiovascular disorders (R^2 change = .01, t = 2.72, p = .02). The other variables did not contribute significantly. For the 1999 recertification examination, similar results were obtained. The critical care medicine certifying examination again entered first (R^2 change = .14, t = 7.74, p < .001), followed once more by the interaction of the frequency and complexity of patients with cardiovascular disorders (R^2 change = .04, t = 3.98, p < .001). Again, the other variables did not contribute significantly.

From these results we can infer that if there were critical care physicians who spent all of their time (100%) treating patients with complex cardiovascular problems they could be expected to perform 3.5 (1997) to 4 (1999) points better on cardiovascular disease items than would those who did not see any cardiovascular problems. There were not many physicians in this sample who spent all of their time treating such patients, but this constitutes a difference of 1.1 to 1.7 standard deviations.

In predicting the pulmonary disorder subscore on the 1997 recertification examination, the critical care medicine certifying examination entered first (R^2 change = .12, t = 8.20, p < .001), followed by frequency of patients with pulmonary disorders (R^2 change = .02, t = 2.57, p < .001). The other variables did not contribute significantly. For the 1999 recertification examination, the critical care medicine certifying examination also entered first (R^2 change = .23, t = 10.20, p < .001), followed by the complexity of patient problems (R^2 change = .08, t = 5.16, p < .001), and frequency of patients with pulmonary disorders (R^2 change = .01, t = 2.58, p = .01). Again, the other variables did not contribute significantly.

From these results we can infer that if there were critical care physicians who spent all of their time treating patients with complex pulmonary problems they could be expected to perform .7 (1997) to 5.2 (1999) points better on pulmonary disease items than would those who did not see any pulmonary problems. There were not many physicians in this sample who spent all of their time treating such patients, but this constitutes a difference of between .4 to 1.6 standard deviations.

Discussion

The purpose of this study was to extend previous work by exploring the relationship between test performance and the nature of practice. Physicians who were recertifying in critical care medicine in 1997 and 1999 supplied information about the amounts of time they had spent in the care of patients with cardiovascular and pulmonary problems and the complexity of the problems they saw. These practice data were compared with performances on the relevant items from the examination.

For cardiovascular diseases, the interaction between volume and complexity had a significant relationship with test scores for both years of the study, even after controlling for previous examination performance. For pulmonary diseases, only volume was a significant predictor in 1997, but both volume and complexity were significant in 1999. The magnitude of the effects was noteworthy, ranging from .4 to 1.7 standard deviations.

These results should be interpreted with care because this study has several limitations. First, the number of questions in each content area was relatively small, and this attenuated the correlations that were reported. Second, the estimates of time and complexity

were based on self-reported data, and a number of physicians made errors in filling out the form. Patients' records would clearly be a more accurate and less biased source of these data. Third, critical care medicine is a relatively new discipline and the vast majority of diplomates are certified in pulmonary disease. Therefore, these results may not generalize to a more homogeneous, less cross-disciplinary field.

Despite these limitations, the results of this study replicate previous work indicating that cognitive examination test scores are associated with patient volume, and by implication from other studies, with patient outcomes. The study also found that there is a relationship between scores and the complexity of problems physicians see in practice. This finding bears more investigation, but it seems sensible that a practice that includes the challenge of treating many complex patients should lead to more knowledge and better judgment on the part of the physician.

These findings, taken together with previous work, suggest that performance on a cognitive examination is related to performance in practice. Of course, this type of examination is not a substitute for rigorous evaluation of practice outcomes, nor is it broad enough to include important aspects of competence such as communication skills and professionalism. Nevertheless, until better measures are available for high-stakes use, the cognitive examination is a reasonable alternative.[14] When such measures become available, there will still be a place for cognitive assessment of new developments in medicine and for patients' problems that are important but infrequently encountered in practice.

The American Board of Internal Medicine supported this research but it does not necessarily reflect its views. Correspondence: John J. Norcini, PhD, Institute for Clinical Evaluation, 510 Walnut Street, Suite 1700, Philadelphia, PA 19106-3699.

References

1. Grambling A. Health plans want to know: are you certified? Managed Care. 1994; May:39–41.
2. Aring CD. Recertification. In: Lloyd JS, Langsley DG (eds). Recertification for Medical Specialists. Evanston, IL: American Board of Medical Specialties, 1987.
3. Showstack JA, Rosenfeld KE, Garnick DW, Luft HS, Schaffarzick RW, Fowles J. Association of volume with outcome of coronary artery bypass graft surgery: scheduled vs nonscheduled operations. JAMA. 1987;257:785–9.
4. Steel K, Norcini JJ, Brummel-Smith K, Erwin D, Markson L. The first certifying examination in geriatric medicine. J Am Geriatr Soc. 1989;37:1188–91.
5. Shea JA, Norcini JJ, Baranowski RA, Langdon LO, Popp RL. A comparison of video and print formats in the assessment of skill in interpreting cardiovascular motion studies. Eval Health Prof. 1992;15:325–40.
6. Norcini JJ, Lipner RS. Recertification: is there a link between take-home and proctored examinations? Acad Med. 1999:74(10 suppl):S27–S30.
7. Tamblyn R, Abrahamowicz M, Brailovsky C, et al. Positive association between licensing examination scores and selected aspects of resource use and quality of care in primary care practice. JAMA. 1998;280:989–96.
8. Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. Ann Intern Med. 1989;110:719–26.
9. Norcini JJ, Kimball HR, Lipner RS. Certification and specialization: do they matter in the outcome of acute myocardial infarction? Acad Med. (in press).
10. Kelly JV, Hellinger FJ. Physician and hospital factors associated with mortality of surgical patients. Med Care. 1986;24:785–800.
11. Rutledge R, Oller DW, Meyer AA, Johnson GJ. A statewide, population-based time-series analysis of the outcome of ruptured abdominal aortic aneurysm. Ann Surg. 1996;223:492–502.
12. Haas JS, Orav EJ, Goldman L. The relationship between physicians' qualifications and experience and the adequacy of prenatal care and low birthweight. Am J Public Health. 1995;85:1087–91.
13. Norcini JJ, Grosso LJ, Shea JA, Webster GD. The relationship between features of residency training and ABIM certifying examination performance. J Gen Intern Med. 1987;2:330–6.
14. Norcini JJ. Recertification in the United States. BMJ. 1999;319:1183–5.

81

Validity of Faculty Ratings of Students' Clinical Competence in Core Clerkships in Relation to Scores on Licensing Examinations and Supervisors' Ratings in Residency

CLARA A. CALLAHAN, JAMES B. ERDMANN, MOHAMMADREZA HOJAT, J. JON VELOSKI, SUSAN RATTNER, THOMAS J. N' SCA, and JOSEPH S. GONNELLA

Connections between assessment measures in medical school, residency, and practice need t be studied in order to ascertain the validity of such assessments in the continuum of medical education and physician training.[1,2] Assuring the validity of students' clinical competence ratings is especially important because these assessments are among the major components of the dean's letter of evaluation and, as such, are used in the ranking of candidates for residency programs.

Medical schools expend considerable time and effort in preparing a dean's letter for each of their graduating students. It is based largely on the faculty's assessment of the student's academic and clinical performance. It should be one of the most important attachments to students' applications for graduate medical education. Despite this, residency directors may not attach much importance to the dean's letter,[3] in part, perhaps, because they are uncertain that the information contained within it is valid for predicting performance during residency.

Previous surveys have indicated that academic criteria such as U.S. Medical Licensing Examinations (USMLE) scores, membership in Alpha Omega Alpha (AOA), the medical honor society, and class rank[4,5] were rated highly as selection variables by residency directors. More recently, performance during clinical clerkships has been cited as an important factor,[3,6] particularly in the specialty for which the student is applying, and especially for the most competitive residencies.[7] It is thus increasingly important to confirm the validity of clerkship evaluations to assure the credibility of the dean's letter as a predictor of postgraduate performance.

The dean's letters of evaluation from Jefferson Medical College include a broad range of information (USMLE Step 1 score, second- and third-year class ranks, histogram of third-year written examination grades, clinical ratings, and excerpts from the narrative evaluations from the third-year clerkships). We have previously documented the validity of a calculated medical school class rank in predicting postgraduate performance.[8,9]

The purpose of this study was to examine the validity of faculty ratings of students' clinical competences in six core clinical clerkships in relation to the students' subsequent performances on medical licensing examinations and to program directors' ratings of clinical performance in the first year of residency.

Method

Study participants were 2,158 students at Jefferson Medical College who graduated between 1989 and 1998. Faculty ratings of stduents' clinical competences in core clerkships in the third year of medical school, scores on lice sing examinations, and residency program directors' ratings of c..nical competence were retrieved from the database of the Jefferson Longitudinal Study of Medical Education.[10]

The predictors (independent variables) included faculty ratings of students' clinical competences in six core c lerkships (family medicine, internal medicine, obstetrics–gynecology, pediatrics, psychiatry, and surgery). These global ratings are part of a detailed assessment form that is completed by the clerkship coordinators at each site. The global ratings of clinical competence in each clerkship were assigned on a five-point scale currently designated as 5 = "high honors," 4 = "excellent," 3 = "good," 2 = "marginal," and 1 = "incomplete" or "failure."

The criterion measures (dependent variables) included scores on USMLE Steps 2 and 3 and postgraduate clinical competence ratings for graduates who had given written permission for follow up (about 75% of the graduating seniors). These ratings were assigned by directors of the residency programs near the end of the first year, using a 33-item rating form. This form measures three areas of clinical competence: "data gathering and processing skills" (16 items), "interpersonal skills and attitudes" (ten items), and "socioeconomic aspects of patient care" (seven items). Each item was rated on a four-point Likert scale, and ratings were averaged within the three competence areas. Data have been reported in support of the measurement properties of this rating form, including construct validity (factor structure), the internal consistency aspect of reliability, and the criterion-related validity of the form.[11,12]

Scores on the USMLE Step 1 were also used to adjust the outcomes for performance differences on this examination. Bivariate correlations and multiple regression analyses were used to examine the associations between ratings in medical school clerkships and the criteria.

Results

The bivariate correlations reported in Table 1 are all statistically significant ($p < .01$). The highest correlations of .29 and .20 for clerkship ratings and USMLE scores were found between the internal medicine clerkship and Steps 2 and 3, respectively. The lowest correlations of .17 and .11 were observed for the psychiatry clerkship and Step 2 scores and for rhe surgery clerkship and Step 3 scores, respectively. Larger correlations were obtained for the internal medicine, family medicine, pediatrics, and obstetrics–gynecology clerkships than for the psychiatry and surgery clerkships.

The results of multiple regression analysis indicated that the shared variance between clerkship ratings and Step 2 scores was 14% ($R^2 = .14$). The overlap was 7% for Step 3 scores, 12% for postgraduate ratings in data gathering and processing skills, 11% for ratings in interpersonal skills and attitudes, and 9% for ratings in the socioeconomic aspects of patient care. Each of these relationships was statistically significant ($p < .01$).

Inspection of the standardized regression coefficients, or beta weights, reported in Table 1 indicate that in a multivariate statistical model, competence ratings given in family medicine, internal medicine, and pediatrics clerkships contributed significantly and consistently to the prediction of all five criterion measures ($p < .01$). The magnitudes of the standardized regression coefficients indicate that among these clerkships, ratings in the internal medicine clerkship had the largest unique contribution in predicting three of the five criterion measures.

Ratings in the psychiatry clerkship contributed to the prediction of Steps 2 and 3 in the multivariate model ($p < .05$), hut did not predict ratings of postgraduate clinical competence. Ratings in the surgery clerkship had a unique contribution to prediction of Step 2, and to ratings for data-gathering and processing skills and interpersonal skills and attitudes.

Additional analyses examined the total number of high-honors

82

TABLE 1. Summary Results of Correlational Analyses of Third-year Students' Clinical Competence Ratings in Six Core Clerkships and the Students' Scores on USMLE Steps 2 and 3 and their Postgraduate Clinical Competence Ratings*

	USMLE				Postgraduate Clinical Competence					
	Step 2		Step 3		Data Gathering†		Interpersonal‡		Socioeconomic§	
Clerkship	(r)	β	(r)	β	(r)	β	(r)	β	(r)	β
Family medicine	(.21)	.11¶	(.18)	.08¶	(.23)	.13¶	(.18)	.09¶	(.21)	.13¶
Internal medicine	(.29)	.19¶	(.20)	.12¶	(.27)	.15¶	(.22)	.11¶	(.22)	.10¶
Obstetrics–gynecology	(.20)	.08¶	(.11)	.00	(.20)	.11¶	(.22)	.13¶	(.18)	.10¶
Pediatrics	(.26)	.11¶	(.19)	.12¶	(.23)	.10¶	(.23)	.12¶	(.20)	.08¶
Psychiatry	(.17)	.06**	(.14)	.07**	(.10)	.01	(.09)	.01	(.10)	.02
Surgery	(.22)	.08¶	(.11)	.01	(.18)	.07**	(.17)	.08¶	(.15)	.04
Multiple R	.38¶		.27¶		.35¶		.33¶		.30¶	

*The total sample included 2,158 graduates of Jefferson Medical College between 1989 and 1998. Bivariate correlations are shown in parentheses. Standardized regression coefficients (beta weights) are shown outside parentheses. All bivariate correlations are statistically significant ($p < .01$).
†Competence ratings of postgraduate clinical skills in "data-gathering and processing."
‡Competence ratings of postgraduate clinical skills in "interpersonal skills and attitudes."
§Competence ratings of postgraduate clinical skills in "socioeconomic aspects of patient care."
¶$p < .01$.
**$p < .05$.

ratings earned by each student across the six clerkships. We classified the numbers of high-honors ratings, which ranged from 0 to 6, into the following three categories: 0 (48% of the sample), 1–3 (48% of the sample), and 4–6 (4% of the sample).

We examined the willingness of the residency program directors to offer further residency training to each resident at the end of the first postgraduate year in relationship to the number of high honors. Further residency, which is usually offered only to those who solidly meet the first-year training standards, was offered to all but 66 (5%) of the 1,401 graduates for whom data were available. We found that the proportion of graduates who would not be offered further training was the highest (6%) among those with no high-honors rating in any clerkship, followed by those with one to three high-honors ratings (3%). All of the graduates with between four and six high-honors ratings were offered further training. The association between the number of high-honors ratings and the offer of further residency training was statistically significant ($\chi^2_{(2)} = 9.4$, $p < .01$).

We conducted additional analyses by adding Step 1 scores to the multiple regression models in predicting the five criterion measures reported in Table 1 to statistically adjust for differences in Step 1 scores. After adjustment, the competence ratings in internal medicine, family medicine, and pediatrics significantly predicted Step 2 scores; and competence ratings in family medicine and pediatrics significantly predicted Step 3 scores. The statistical control of Step 1 scores did not change the pattern of findings in multivariate regression analysis in which the ratings of competence in the core clerkships were the predictor and ratings of the three postgraduate clinical competence areas of "data-gathering and processing skills," "interpersonal skills and attitudes," and "socioeconomic aspects of patient care" were the criterion measures.

Discussion

The present study examined the validity of clinical competence evaluations assigned by medical school faculty, which are often reported in dean's letters of evaluation. Our findings suggest that faculty ratings are valid and are useful in predicting performances on medical licensing examinations and clinical competence ratings in residency. Although the faculty ratings assigned in the internal medicine, family medicine, and pediatrics clerkships yielded stronger associations with the criterion measures than did those in the psychiatry and surgery clerkships, the number of high-honors ratings that a student earned in all six clerkships was found to have a significant association with whether or not further training was offered to the graduate at the end of the first year of residency.

It should be noted that although the correlation coefficients were all statistically significant, they were not large. All fell in the range of small to moderate effect size estimates described by Cohen.[13] However modest in magnitude, the consistency of the results provides credible evidence in support of the validity of the ratings.

Conclusions and Implications

Medical schools want to help each of their graduates to obtain the best residency position commensurate with his or her qualifications. However, most faculty realize that it is shortsighted to prepare a dean's letter that misrepresents a student's medical school record or excludes relevant observations of the student's performance. Obfuscation is counterproductive.[14] We found that the clerkship ratings for internal medicine, family medicine, pediatrics, and obstetrics–gynecology were significantly correlated with criterion measures. These evaluations were significant predictors of performance in postgraduate training. Likewise, our findings indicate that the high-honors ratings of competence in core clerkships were significantly associated with residency program directors' decisions to offer further residency training.

The largest correlations were obtained for ratings in the internal medicine clerkship. This could be due to the fact that our students spend 12 weeks on this, and only six weeks on the others. This expanded time in the internal medicine clerkship allows for more observations and broader evaluations by a larger number of faculty and residents that could contribute to an increased overlap between this clerkship's ratings and the criterion measures.

The Association of American Medical Colleges recommended in 1989 that the dean's letter be described as a letter of evaluation rather than as a letter of recommendation.[15] Many have followed this recommendation. Studies in a variety of settings have confirmed that superior performance in medical school does predict performance beyond medical school.[1,10,16] Our results should not only increase the confidence of the medical school faculty with respect to their evaluations, but also reassure residency selection committees about the validity of evaluations in dean's letters as predictors of clinical competence beyond medical school. Every

83

medical school should be committed to provide empirical support for the validity of information in its dean's letters of evaluation.

Correspondence: Clara Callahan, MD, Admissions Office, Jefferson Medical College, Philadelphia, PA 19107-5833; e-mail: ⟨clara.callahan@mail.tju.edu⟩.

References

1. Gonnella JS, Hojat M, Erdmann JB, Veloski JJ (eds). Assessment Measures in Medical School, Residency, and Practice. New York: Springer, 1993.
2. Campos-Outcalt D, Witzke DB, Fulginiti JV. Correlations of family medicine clerkship evaluations with scores on standard measures of academic achievement. Fam Med. 1994;26:85–8.
3. Wagoner NE, Suriano JA. Recommendations for changing the residency selection process based on a survey of program directors. Acad Med. 1992;67:459–65.
4. Wagoner NE, Grey G. Report of a survey of program directors regarding selection factors in graduate medical education. J Med Educ. 1979;54:445–52.
5. Wagoner NE, Suriano JR, Stoner JA. Factors used by program directors to select residents. J Med Educ. 1986;61:10–21.
6. Villanueva AM, Kaye D, Abdalhak SS, Morahan PS. Comparing selection criteria of residency directors and physicians' employers. Acad Med. 1995;70:261–71.
7. Wagoner NE, Suriano JR. Program directors' responses to a survey on variables used to select residents in a time of change. Acad Med. 1999;74:51–8.
8. Blacklow RS, Goepp CE, Hojat M. Class ranking models for dean's letters and their psychometric evaluation. Acad Med. 1991;66(9 suppl):S10–S12.
9. Blacklow RS, Goepp CE, Hojat M. Further psychometric evaluations of a class-ranking model as a predictor of graduates' clinical competence in the first year of residency. Acad Med. 1993;68:295–7.
10. Hojat M, Gonnella GS, Veloski JJ, Erdmann JB. Jefferson Medical College's longitudinal study: a prototype of assessment of changes. Education for Health. 1996; 9:99–113.
11. Hojat M, Veloski JJ, Borenstein BD. Components of clinical competence ratings: an empirical approach. Educ Psychol Meas. 1986;46:761–9.
12. Hojat M, Borenstein BD, Veloski JJ. Cognitive and noncognitive factors in predicting the clinical performance of medical school graduates. J Med Educ. 1988; 63:323–5.
13. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Erlbaum, 1987.
14. Edmund M, Roberson M, Hasan N. The dishonest dean's letter: an analysis of 532 dean's letters from 99 US medical schools. Acad Med. 1999;74:1033–5.
15. Association of American Medical Colleges. Report of the Ad Hoc Committee on Dean's Letters. In: A Guide to the Preparation of Medical School Dean's Letter. Washington, DC: AAMC, 1989.
16. Hojat M, Gonnella JS, Erdmann JB, Veloski JJ. The fate of medical students with different levels of knowledge: are the basic medical sciences relevant to physician competence? Adv Health Sci Educ. 1997;1:197–207.

84

Do Students' Attitudes during Preclinical Years Predict Their Humanism as Clerkship Students?

JOHN C. ROGERS and LOUISA COUTTS

There is an increased awareness of the importance of humanism in the medical school curriculum.[1] Most of the early work concerned teaching methods[2-6] and measurement of humanism.[7-14] Contemporary work reinforces the critical role of empathy in humanism[15] but also broadens the concept to include other values, qualities, and behaviors: authenticity, compassion, fidelity, integrity, respect, spirituality, and virtue.[16-21] To distinguish it from humanism, professionalism is characterized as accounrability, altruism, commitment to excellence, duty and commitment to service, and honor and respect for others.[18] The project we report here began over eight years ago, so compassion, empathy, respect, and considerate biopsychosocial interactions are the conceptual cornerstones of the operational definition we used in this work; we therefore defined the humanistic physician as one who

1. respects the patient's viewpoints and considers his or her opinions when determining health care decisions,
2. attends to the psychological well-being of the patienr,
3. regards the patient as a unique individual,
4. treats the patient in the context of his or her family and social and physical environment,
5. possesses good communication and listening skills,
6. engenders trust and confidence,
7. demonstrates warmth and compassion, and
8. is empathetic.[22]

Despite the considerable attention to humanism in medical education, little is known about predictors of humanism in students. Knowledge about predictors of humanism could foster the evaluation and design of curricular innovations by identifying attitudes that may be affected by educational interventions. The objective of our work was to identify potentially modifiable attitudes that are associated with students' humanistic performance during clinical clerkships. These predictors may be important outcome measures for the curricular interventions, and may help assess the need for additional innovations.

Methods

Between 1992 and 1995, we had students complete attitude questionnaires during a second-year required preclinical course and again during a required third-year clerkship in family and community medicine. The third-year students also completed a clinical performance examination (CPX) where standardized patients (SPs) rated the students' humanism.

We administered four previously developed attitude questionnaires: (1) Physician Belief Scale,[23] (2) Physician Reactions to Uncertainty,[24] (3) Risk in Clinical Practice,[25] and (4) Decision Making Style.[26] We had students complete these instruments to provide students with feedback on their attitudes that might affect their clinical behaviors. We gave each student a confidential report with his or her score for each scale, the class's average score and range for each scale, and the normative averages and ranges from the instrument-development samples. The third-year students' report included their second-year scores, so each student could reflect on any changes in attitude scores after experience in clinical rotations. We considered the concepts measured by the instruments to be important to general clinical performance, and not specifically or solely to be predictors of humanism. Similarly, we chose an available measure of humanism that could be completed by SPs to give students feedback on this particular aspect of clinical performance.

The Physician Belief Scale is a 32-item questionnaire about ways physicians who adopt a biopsychosocial approach to patient care differ from physicians who do not. Responses are recorded on a five-point Likert-type scale ranging from 1 = disagree to 5 = agree. The scale scores range from 32 (maximum degree of psychosocial orientation) to 160 (minimum psychosocial orientation). The authors of the instrument determined the internal consistency of the scale by Kuder–Richardson formula 20, which measures the extent to which the items reflect a single underlying construct. This unidimensional scale is highly internally consistent, with r = 0.88. The mean score of the 180 family physicians, psychiatrists, and internists in the development sample was 74.3 (SD = 13.7), and that of the 99 family physicians, psychiatrists, internists, and pediatricians in the validation sample was 72.1 (SD = 13.0).[23] Eight faculty family physicians in the Department of Family and Community Medicine had a mean score of 58.4 (SD = 10.1).

The Physician Reactions to Uncertainty scale measures physicians' affective reactions to uncertainty, which "seem to be a significant, yer overlooked, dimension of patient care decisions and variations in practice patterns."[24] This scale consists of two subscales derived from factor analysis: (1) Stress from Uncertainty subscale (13 items) and (2) Reluctance to Disclose Uncertainty to Others subscale (nine items). Both subscales use a six-point Likert response scale from 1 = strongly agree to 6 = strongly disagree, with many items reverse-scored to prevent response-set bias. The Stress from Uncertainty subscale ranges from 13 to 78 (the higher the score the greater the stress) and has a Cronbach alpha of .90, indicating excellent inrernal consisrency. The mean score for 428 family physicians, general practitioners, general internists, medical subspecialists, and surgeons in the development sample was 44 (SD = 11). The Reluctance to Disclose Uncertainty to Others subscale ranges from 9 to 39 (the higher the score the greater the reluctance to disclose uncertainty) and has a Cronbach alpha of .75, indicating acceptable internal consistency. The mean score for the 428 physicians was 23 (SD = 6).[24]

The Risk in Clinical Practice questionnaire measures physicians' general self-perceptions of levels of risk aversion/risk seeking and risk attitudes in financial, physical well-being, social, and ethical domains. One nine-point Likert scale item is used for each domain (1 = avoid risk/danger to 9 = seek risk/danger). In a study of laboratory usage, 12 family physicians completed the questionnaire, producing a mean score for each item: gambling for money (3.6, SD = 1.8), physical danger (4.6, SD = 1.4), hurting people's feelings (reverse-scored, 5.3, SD = 1.9), professional norms (reverse-scored, 6.2, SD = 1.7), risks to self (4.7, SD = 1.4), and risks to patients (3.8, SD = 1.6). Significant positive rank-order correlations were observed between: gamble and norms, risk self and gamble, risk self and danger, and risk self and norms.[25] In the present study, a student's aggregate risk score was the proportion of the maximum possible score. A measure of internal consistency for this scale is not available.

The Decision-making Style is a 32-item questionnaire that measures the extent to which a person is intuitive or analytic in making decisions. Each item presents a forced choice between two alternatives. Scores range from 0 to 32, with low scores indicating an-

85

TABLE 1. Medical Students' Attitudes about Psychosocial Aspects of Care, Uncertainty, Risk, and Decision-making Style—Comparison of Students' Attitude Scores during their Preclinical and Clinical Years with Attitude Scores of Physicians in Instrument-development Samples and Correlation of Students' Attitude Scores with Humanism Ratings by Standardized Patients during a Third-year Clerkship Clinical Performance Examination, Baylor College of Medicine, 1992–1995*

	Attitude-scale Scores			Correlation of Attitudes with Humanism	
Attitude Instrument	Physicians in Instrument Development Samples	Students' Preclinical Scores (n = 299)	Students' Clinical Scores (n = 366)	Students' Preclinical Scores (n = 299)	Students' Clinical Scores (n = 366)
Physician Belief Scale (range 32–160); higher score means lower belief in psychosocial aspects of care	73.2	72.2	73.5	−.252 $p < .001$	−.222 $p < .001$
Physician Reactions to Uncertainty (range 13–78); higher score means more stress from uncertainty	44	48	46	−.049 $p = .230$	−.010 $p = .424$
Physician Reactions to Uncertainty (range 9–39); higher score means greater reluctance to disclose uncertainty to others	23	29	28	−.042 $p = .265$	−.115 $p = .014$
Risk in Clinical Practice (range 1–9); higher score means more risk seeking	.52	.49	.50	.090 $p = .087$	−.016 $p = .378$
Decision-making Style (range 0–32); higher score means more intuitive than analytic approach to decisions		17.2	17.1	.174 $p = .004$	.021 $p = .347$

*We administered standardized attitude instruments to medical students during their second preclinical year and again during their third-year clerkship in family and community medicine. We also administered a clinical performance examination during the clerkship, where standardized patients rated the students' humanism using a validated scale. We correlated the standardized patients' humanism ratings with the students' preclinical and clinical attitude scores.

alytic approaches and high scores indicating intuitive approaches to decisions and problems.[26] A measure of internal consistency for this scale is not available.

Standardized patients (SPs) rated students' performances during a CPX on interview style, history taking, physical examination, management negotiation, and patient education. The SPs completed an encounter checklist and a separate humanism questionnaire for each student in each encounter. Interstation exercises assessed the students' differential diagnoses, management plans, and identification of ethical principles. The CPXs were conducted during the second and fifth weeks of the six-week rotation, and each CPX had a minimum of five cases. Students completed between ten and 13 patient cases in the CPX, with ten minutes for the SP activity and five minutes for the interstation activity. The CPX interstation reliability alpha coefficients ranged between 0.43 and 0.56. These moderate scores are comparable to the reliabilities reported in the literature for CPXs that used similar numbers of cases and tested similarly wide varieties of clinical skills, such as those encountered in family medicine.[27,28] A student had to complete at least one full CPX with a minimum of five cases to be included in this study, in order to have multiple SP ratings of humanism and hopefully a stable measure for each student.

For humanism, the SPs rated each student at each CPX station using the eight-item, abbreviated Humanism Scale developed by Hauck et al.[22] The full Humanism Scale is a 24-item questionnaire assessing whether a physician has a "sensitive, non-humiliating, and empathetic way of helping [a patient] deal with some problem or need" and correlates highly with patients' satisfaction with physician-related aspects of care.[22] The full scale includes the eight components listed in the first paragraph of this article, with three items each. The eight-item scale has the following items (one per component):

1. This doctor seems to take a personal interest in me.
2. Even when my problem is small, this doctor is concerned.
3. I have confidence in this doctor's decisions.
4. This doctor respects my beliefs.
5. I would talk to this doctor if something were troubling me.
6. This doctor takes an interest in my home life.
7. This doctor is easy to talk to.
8. This doctor seems to know what I am going through when I tell him/her about a problem.

In the development study, responses were recorded as an "x" on a line between strongly disagree and strongly agree. The response point was measured with a ruler and converted to a percentage of the total line (1 to 99 for each item). The scale score was the mean for the 24 items. The development sample of 185 patients produced humanism scores ranging from 16 to 99 (1 to 99 is the possible range), with a mean of 75. Cronbach alpha (reliability coefficient) for the 24 items was .95; the eight-item scale with one item per component had a coefficient of .93. We used a seven-point Likert response scale (1 = strongly disagree, 7 = strongly agree) for the eight-item scale, which had a Cronbach alpha of .96. To adjust for variability among SPs on these ratings, we normalized each SP's humanism scores to the average for all SPs.

We used SPSS for Windows® for data analysis to produce Pearson correlation coefficients.

Results

The students' scores for the attitude scales were quite similar to those of the development samples of experienced clinicians for the Physician Belief Scale and the Risk in Clinical Practice Scale. The students appeared to have more stress from uncertainty and reluctance to disclose uncertainty to others than the experienced clinicians in the development sample (Table 1). The standardized patients' ratings of the students' humanism also were quite similar to those of the development sample; when both scale scores are converted to a proportion of the maximum possible score (mean/maximum): development sample 75/99 = .76 and students 42/56 = .75.

The students' preclinical Physician Belief Scale scores were significantly inversely correlated with clerkship humanism (Table 1).

Students who rated the psychosocial aspects of medicine lower (higher Physician Belief Scale score) had lower humanism scores. This relationship persisted for students' psychosocial beliefs during the clerkship. Students' preclinical decision-making style was directly related to humanism, with more intuitive students having higher scores on the humanism scales, but this relationship was not stable into the clinical year. The students' preclinical Reluctance to Disclose Uncertainty to Others scores were not significantly related to humanism, but their clinical ratings were inversely related, indicating higher levels of humanism in students less reluctant to disclose their uncertainty. The students' Stress from Uncertainty and their Risk in Clinical Practice scores were not related to humanism.

Discussion

These students' attitudes about biopsychosocial aspects of care and risk in clinical practice appear similar to those of experienced clinicians, but students may be more stressed by uncertainty and reluctant to disclose it to others than experienced clinicians. The standardized patients' ratings of the students' humanism were similar to the ratings the patients gave practicing physicians. These results lend some support to the legitimacy of using these instruments with students.

The students' preclinical attitudes toward the biopsychosocial aspects of medical care are a potential predictor for their humanism on clinical rotations. The Physician Belief Scale was developed as a self-report instrument for practicing physicians, but may be useful as a predictive tool for students. Attitudes about uncertainty, risk, and decision making do not appear to be consistently related to humanism.

These results are sensible considering the eight items that compose the Humanism Scale. The consistent relationship with the Physician Belief Scale indicates that students who do not value highly the biopsychosocial aspects of care are not be able to display the interest, concern, and respect necessary for high ratings of humanism by standardized patients. The inconsistent relationship for the Reluctance to Disclose Uncertainty to Others and Decision-making Style suggests that students comfortable with sharing their uncertainty or those preferring an intuitive decision style may be perceived by standardized patients as open about decisions, easy to talk to, or knowing what the standardized patients are going through. The concepts of Stress from Uncertainty and Risk in Clinical Practice aren't as obviously related to the components of humanism and were not associated with humanism in this study. The concepts of uncertainty, risk, and decision making may still be important for clinical performance in general but not for humanism in particular. On the other hand, the concepts inherent in biopsychosocial attitudes seem to be related to clinical performance of the concepts of humanism we measured in this study: compassion, respect, empathy, and especially considerate biopsychosocial interactions.

The limitations of this work include questions about both the internal and external validity. It is unclear how the number of cases completed by students could have affected the stability of the humanism measure or influenced the results. The potentially problematic reliability of the Risk in Clinical Practice and Decision-making Style scales could have contributed to the failure to detect relationships between those scales and humanism. Even the significant correlations we did observe are small, and account for little of the variability in humanism, so their statistical significance simply may be due to the sample size. The generalizability of these results is limited, since our study involved only one institution with approximately two class cohorts of students. Studies in other medical schools with additional cohorts of students would determine whether these apparent associations are stable.

Further instrument development may improve the ability of attitude measures to predict later humanistic behaviors. The Physician Belief Scale seems to measure well attitudes associated with considerate biopsychosocial interactions. Since contemporary definitions of humanism include so many concepts (compassion, respect, empathy, integrity, fidelity, authenticity, spirituality, and virtue), we may need one comprehensive instrument, or several separate instruments, to measure the attitudes corresponding to the many facets of humanism. Demographic variables that may predict humanism, such as gender, could be included in predictive models, but the fundamental purpose of the work we report here is the identification of potentially modifiable predictors. Including demographic variables may improve the explanatory power of a multivariate model, but will it lead to an admission policy of preferentially selecting students with the unchangeable demographic variables positively associated with humanism? Or should we concentrate our efforts on curricula and the attitudes that we may be able to influence?

Curricula in many medical schools have courses emphasizing the physician–patient relationship and problem-based learning, which may influence the students' preclinical attitudes found in our work to be associated with humanism. Future work may reveal relationships between other attitudes, which are conceptually related to humanism, and students' humanistic behaviors in clinical rotations. Measurements of both attitudes and behaviors may be important outcome measures for curricular interventions, and may help assess the need for additional innovations.

Correspondence: Dr. John C. Rogers, Baylor Family Medicine, 5510 Greenbriar, Houston, TX 77005; e-mail: ⟨jrogers@bcm.tmc.edu⟩.

References

1. Wilkes MS, Slavin SJ, Usatine R, Wilkes M. Doctoring—a longitudinal generalist curriculum. Acad Med. 1994;69:191–3.
2. Pellegrino E. Educating the humanistic physician. JAMA. 1974;227:1288–94.
3. Pence G. Can compassion be taught? J Med Ethics. 1983;9:189–91.
4. Spiro H. What is empathy and can it be taught? Ann Intern Med. 1992;116:843–6.
5. Hahn R. A method for teaching human values in clinical clerkships through group discussion. Teach Learn Med. 1991;3:143–50.
6. Engelberg J. A program of integrative humanistic study for medical students. Acad Med. 1992;67:455.
7. Abbott L. A study of humanism in family physicians. J Fam Pract. 1983;16:1141–6.
8. Beckman H, Frankel R, Kihm J, Kulesza G, Geheb M. Measurement and improvement of humanistic skills in first year trainees. J Gen Intern Med. 1990;5:42–5.
9. Butterfield PS, Mazzaferri EL. A new rating form for use by nurses in assessing resident's humanistic behavior. J Gen Intern Med. 1991;6:155–61.
10. Simmons JM, Robie PW, Kendrick SB, Schumacher S, Roberge LP. Residents' use of humanistic skills. Am J Med Sci. 1992;303:227–32.
11. Ramsey PG, Wenrich M. Evaluation of humanistic qualities and communication skills. J Gen Intern Med. 1993;8:164.
12. Weaver MJ, Ow CL, Walker DJ, Degenhardt EF. A questionnaire for patients' evaluations of their physicians' humanistic behaviors. J Gen Intern Med. 1993;8:135–9.
13. McLeod PJ, Tamblyn R, Benaroya S, Snell L. Faculty ratings of resident humanism predict satisfaction ratings in ambulatory medical clinics. J Gen Intern Med. 1994. 9:321–6.
14. Woolliscroft JO, Howell JD, Patel BP, Swanson DB. The humanistic qualities of internal medicine residents assessed by patients, physicians, supervisors, and nurses. Acad Med. 1994;69:216–24
15. Marcus ER. Empathy, humanism, and the professionalization process of medical education. Acad Med. 1999;74:1211–5.
16. Kopelman LM. Values and virtues: how should they be taught? Acad Med. 1999;74:1307–10
17. Novack DH, Epstein RM, Paulsen RH. Toward creating physician–healers: fostering medical students' self-awareness, personal growth, and well-being. Acad Med. 1999;74:516–20.
18. Markakis KM, Beckman HB, Suchman AL, Frankel RM. The path to professionalism: cultivating humanistic values and attitudes in residency training. Acad Med. 2000;75:141–50

87

19. Shelton W. Can virtue be taught? Acad Med. 1999;74:671–4.
20. Risdon C, Edey L. Human doctoring: bringing authenticity to our care. Acad Med. 1999;74:896–9.
21. Sulmasy DP. Is medicine a spiritual practice? Acad Med. 1999;74:1002–5.
22. Hauck FR, Zyzanski SJ, Alemagno SA, Medalie JH. Patient perceptions of humanism in physicians: effects on positive health behaviors. Fam Med. 1990;22:447–52.
23. Ashworth CD, Williamson P, Montano D. A scale to measure physician beliefs about psychosocial aspects of patient care. Soc Sci Med. 1984;19:1235–8.
24. Gerrity MS, DeVellis RF, Earp JA. Physicians' reactions to uncertainty in patient care. A new measure and new insights. Med Care. 1990;28:724–36.
25. Holtgrave DR, Lawier F, Spann SJ. Physicians' risk attitudes, laboratory usage, and referral decisions: the case of an academic family practice center. Med Decis Making. 1991;11:125–30.
26. Westcott M. Correlates of intuitive thinking. Psychol Rep. 1963;12:595–613.
27. Petrusa ER, Lackwell TA, Ainsworth MA. Reliability and validity of an objective structured clinical examination for assessing the clinical performance of residents. Arch Intern Med. 1990;150:573–7.
28. Roberts J, Norman G. Reliability and learning from the objective structured clinical examination. Med Educ. 1990;24:219–23.

Early Identification of Students at Risk for Poor Academic Performance in Clinical Clerkships

SCOTT A. FIELDS, CYNTHIA MORRIS, WILLIAM L. TOFFLER, and EDWARD J. KEENAN

Many medical schools have revised, or are in the process of revising, their curricula.[1] The impetus for this curricular change has been dependent on many factors. These factors include grant initiatives emphasizing the development of curricula to promote generalism and the Association of American Medical Colleges' Medical School Objectives Project (MSOP), as well as significant shifts in the health care system, such as the growing influence of managed care. The more innovative curricular revisions to date have included multidisciplinary, integrated courses with longitudinal curricula and early clinical experiences throughout the first two years (the preclinical curriculum).

Oregon Health Sciences University (OHSU) School of Medicine implemented its curriculum revision in 1992.[2] The result of this effort was the restructuring of the first two years of the curriculum from 24 specific discipline-based courses to ten interdisciplinary units. One of the units, the Principles of Clinical Medicine (PCM), is a longitudinal two-year course composed of small-group activities half a day each week and a weekly half-day clinical preceptorship. In addition, there are nine integrated basic science courses in the first two years, and a one-week course, Transition to Clerkship, occurs at the end of the second year. The core clerkships, constituting the entire third year, include medicine, surgery, obstetrics–gynecology, family medicine, psychiatry, pediatrics, and rural primary care. Each of these clerkships is six weeks in duration with the exception of medicine, which occurs in two six-week blocks.

The premise for this study was that early identification of medical students who are at academic risk provides a basis for intervention with individualized remedial programs. Previously, studies have investigated predictors of performance for years one and two of medical school.[3,4] Little has been done to address early identification of students at risk for academic difficulty in the third year of medical school. The hypothesis was that performance in PCM during the predominantly pre-clinical curriculum of the first two years predicts students at risk for academic difficulty in the clinical clerkships. Accordingly, this study analyzed the relationship between parameters of student assessment, including a number of admission, curriculum, and standardized testing criteria, and an accepted standard of graded performance in the third-year core clerkships.

Method

The sample studied was a cohort of students beginning with those who matriculated at OHSU from 1992 to 1995 and who graduated between 1996 and 1999. Student data were available from OHSU databases; no major change in curriculum, grading policy, or calculation of student grade-point averages occurred during these years. In the study, all individual student performance data were treated as confidential.

The primary outcome of this analysis was performance in the core clinical clerkships of the third year curriculum, which serves as a critical component of the residency application process. All courses at OHSU, including clerkships, are graded as honors, near honors, satisfactory, marginal, or fail. Grade-point average (GPA) in year three was used as the outcome, with a GPA of 3.0 representing honors; 2.0 near honors; 1.0 satisfactory, and 0 marginal/failure. After initial analysis as a continuous variable, we identified the lowest quintile of performance in year three (GPA < 2.0).

A number of potential indicators were considered to predict performance in year three. These indicators included cumulative college GPA, separate MCAT scores (Verbal Reasoning, Biological Science, Physical Science, and Writing Sample), year one and year two basic science course performance as a mean percentage examination score, performance in the PCM course, and USMLE Step 1 score. The total MCAT score combined the Verbal Reasoning, Biological Science, and Physical Science scores. For the Writing Sample, the alphabetic score was coded from 4 to 15, with M = 8. Total points for the PCM course were used. In PCM, there are 80 points available for each of six quarters: 20 points for the clinical preceptorship, 10 points for small-group discussion activities, 10 points for patient examination activities, 10 points for an essay, 10 points for written exam, and 20 points for a group objective structured clinical examination (GOSCE).[1]

The first series of analyses were univariate, with all continuous predictor variables correlated with the primary outcome, year three GPA. Subsequently, a parsimonious logistic regression model was fit to predict this outcome using forward selection procedures. The odds of low performance (year three GPA < 2.0) were estimated. Cutoff points for categorizing each continuous predictor variable were based on the lowest quintile of each score or percentage, with latitude for ties. The significance of each predictor variable was assessed using a likelihood-ratio test statistic obtained from a logistic regression model fit to the outcome status.

Results

In total, data for 306 students were available. All data were complete except for one student who had attended a college without grades, seven students who had taken the earlier version of the MCAT, and two students whose USMLE scores were unavailable.

Correlation coefficients were obtained for each performance indicator as compared with the year-three GPA. Of all variables, this outcome was most significantly related to the score in the PCM course ($r = .61$, $p < .001$); year two percentage performance ($r = .54$, $p < .001$); year one percentage performance ($r = .52$, $p < .001$); and USMLE 1 score ($r = .47$, $p < .001$). Year-three GPA was only modestly related to undergraduate GPA ($r = .19$, $p < .05$) and MCAT Writing Sample score ($r = .16$, $p < .05$), and was not related significantly to the total MCAT score, or to the Biological Science, Physical Science, and Verbal Science Subscores. Figure 1 shows the relationship between the year-three GPA and the PCM score.

Prior to logistic regression analysis, the relationship of each variable to performance in the lowest quintile of year-three GPA was analyzed in order to determine the accuracy of prediction. Each was dichotomized by the lowest quintile and compared in a 2×2 table with low year-three GPA. A score in the lowest quintile of PCM (≤ 380) correctly predicted low year three performances of 38 of 68 students (positive predictive value = 56%). Of 238 students who had score above the lowest quintile, 212 (negative predictive value = 89%) had year-three GPAs above the lowest quintile. These values were similar considering performance in the lowest quintile in year two (positive predictive value = 53%, predicting 36 of 68; negative predictive value = 89%, 211 of 238). A USMLE Step 1 score in the lowest quintile (≤ 190) correctly predicted 28 of 68 students who scored in the lowest quintile of year-three GPA (positive predictive value = 41%), whereas a score above the lowest

89

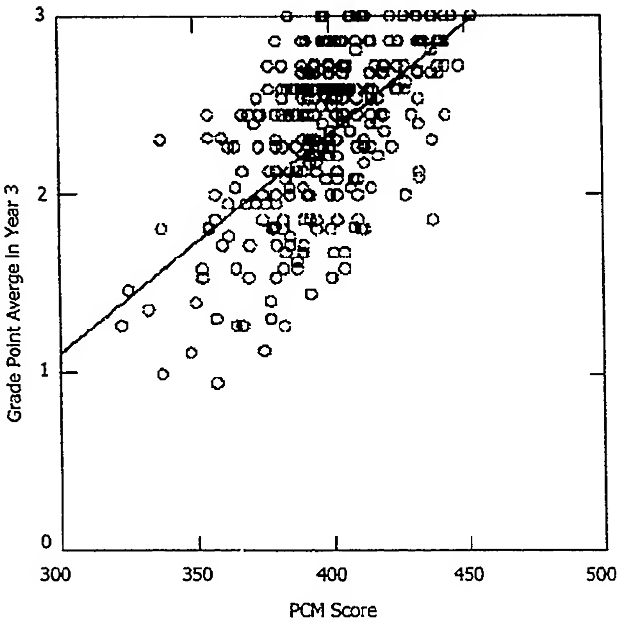

Figure 1. Scatterplot of the relationship between Principles of Clinical Medicine (PCAA) scores and year three grade-point averages at Oregon Health Sciences University School of Medicine, 1992–1999.

quintile predicted 206 of 238 (negative predictive value = 87%). No other variable performed similarly in univariate analysis.

A multivariate logistic regression model significantly predicted low year-three GPA ($p < .001$). (See Table 1.) Overall, performance in the lowest quintile in PCM was associated with a 9.45 times increased risk of performance in the lowest quintile of year-three GPA (95% CI, 4.71–18.98). Similarly, performance in the lowest quintile of year two conferred a 6.39 times risk of low year-three GPA (95% CI, 2.96–13.80). This model also included performance in the lowest quintile of the USMLE Step 1, although this was not significant (relative risk 1.83; 95% CI, 0.84–3.99). Last, performance by quintile of PCM score, after adjustment for USMLE Step 1 score and year-two percentage score, was linearly related to year-three GPAs less than 2.0 ($p < .001$). This confirms that the PCM score has a strong, graded relationship to performance in the clinical clerkships. A student receiving a PCM score in the second lowest quintile was 75% less likely (relative risk 0.25) to perform poorly in the clerkships, as compared with those with scores in the lowest quintile. A PCM score in the highest quintile was associated with a markedly reduced chance of poor performance; in fact, only one student in the highest quintile had a year-three GPA below 2.0.

Discussion

This study may be considered unique to the OHSU curriculum, yet many schools are developing similarly integrated curricula with early clinical experiences. This method of identifying students early in medical school who are at risk for academic and professional difficulties may be generalizable.

In designing this study, we defined the outcome of interest as year-three GPA. As mentioned, this was chosen due to the established connection between clerkship evaluations and residency applications. However, it is important to remember that the validity of third-year clerkship evaluations as an indicator for performance as a physician is unclear.

Our conclusion is that performance in the PCM course is predictive of performance in the core clerkships of the third-year curriculum. Additionally, by identifying students who perform in the lowest quintile of the PCM course, it is possible to identify the students who will be in the lowest quintile of the core third-year clerkships. One explanation of this relationship is that evaluation of performance in the PCM course better assesses the ability to use core knowledge, as well as the evaluation of patient care skills and professional attributes. Consequently, the assessment of student performances in the PCM course may coincide more closely with the approach to evaluation used in the core third-year clerkships. That is, greater emphasis is placed on preceptor and interactive session evaluations, with only a relatively small component related to performance on didactic written examinations. Thus, performance in PCM reflects behavioral and attitudinal factors associated with patient care, in addition to knowledge and skills, in contrast to a singular focus on cognitive performance as is typically the case in the basic science courses.

A concern raised in regard to the PCM course is the "subjective" nature of the performance assessment, particularly in comparison with the "objective" process employed in the basic science courses. This study supports the current evaluation approach in PCM that includes assessments by small-group facilitators and preceptors and GOSCE performance in a fashion that is more consistent with assessment methods used in the third-year clinical clerkships.

Overall, we believe that this information has improved the faculty's confidence in their ability to evaluate student performance. Previously, there were very few students identified as having difficulties prior to the third year, and these students rarely were noted to have professional development issues. The outcome of this study has already influenced the student assessment process in the OHSU School of Medicine. Validation of early concerns about student performance provides faculty with greater confidence in early identification of students who are at academic risk. Confidence in this evaluation system has initiated changes in the medical school's Student Progress Committee's approach to considering at-risk students. A professional development evaluation has been established that is used to identify early concerns regarding professionalism or concerns related to clinical skills or attitudes despite the fact that a student may have successfully passed the courses. Thus, students who are succeeding in the basic science curriculum but who are struggling with clinical integration or who are not demonstrating appropriate professional development may be reviewed by the Student Progress Committee.

Additional factors not considered in the current study, including age, gender, ethnicity, and years between matriculation and graduation from college, may have also contributed to the variations observed in this population. Further analysis is needed to resolve the potential influences of any of these additional factors.

Finally, confidence in identifying students at risk early in the

Table 1. Logistic Regression of Performances in the Lowest Quintiles in the Seven Third-year Core Clinical Clerkships at Oregon Health Sciences University School of Medicine, 1992–1999

	Coefficient	Standard Error	Relative Risk	95% Confidence Interval
Lowest quintile of PCM* score (≤ 380)	2.24	0.36	9.45	4.71, 18.98
Lowest quintile of year two % performance (< 82%)	1.86	0.39	6.39	2.96, 13.80
Lowest quintile of USMLE Step 1 score (≤ 190)	0.60	0.40	1.83	0.84, 3.99

* PCM = Principles of Clinical medicine course.

curriculum provides the opportunity for remediation at a time that is more conducive to improving the long-term success of the student. This creates the need for a process of developing individualized programs to address the specific shortcomings identified. However, an essential aspect of such a process is validation of an early academic warning system, as demonstrated in this study based on assessment during a longitudinal clinical experience in the first two years of medical school. As more medical school curricula now include early clinical experiences, the opportunity exists for confirmation of these findings through multi-site studies.

Correspondence: Dr. Scott A. Fields, Department of Family Medicine, Oregon Health Sciences University, 3181 SW Sam Jackson Park Road, Portland, OR, 97201; e-mail: ⟨safields@ohsu.edu⟩.

References

1. Curry R, Makoul G. The evolution of course in professional skills and perspectives for medical students. Acad Med. 1998;73:10–3.
2. Fields S, Toffler W, Elliot D, Chappelle K. Principles of Clinical Medicine, Oregon Health Sciences University, School of Medicine. Acad Med. 1998;73:25–31.
3. Hall FR, Bailey BA. Correlating students' undergraduate science GPAs, their MCAT scores, and the academic caliber of their undergraduate colleges with their first-year academic performances across five classes at Dartmouth Medical School. Acad Med. 1992;67:121–3.
4. Croen LG, Reichgott M, Spencer RK. A performance-based method for early identification of medical students at risk of developing academic problems. Acad Med. 1991;66:486–8.
5. Elliot DL, Fields SA, Keenen, TL, Jaffe AC, Toffler WL. Use of a GOSCE (Group Objective Structured Examination) with first-year medical students. Acad Med. 1994;69:990–2.

● THOUGHTS ON THINKING

Moderator: Glenn Regehr, PhD

The Under-weighting of Implicitly Generated Diagnoses

KEVIN W. EVA and LEE R. BROOKS

Imagine that a diagnostician is asked to comment on a diagnosis proposed by a colleague. Clearly, to decide whether or not the diagnosis is probably correct, other diagnostic possibilities for that case must be considered. However, the prevalence of confirmation biases recorded in the psychology literature suggests that the proposed diagnosis has some priority over self-generated diagnoses.[1] Considering the proposed diagnosis first might lead to noticing and taking seriously the features consistent with it and evaluating other diagnoses in that light. The current study was designed to address this issue by examining differences in probability ratings and patient management decisions as a function of whether diagnostic alternatives are presented explicitly or are generated by the diagnosticians themselves. Normatively, there should be no difference in the probability assigned to, or the patient management decisions made on the basis of, a diagnostic alternative regardless of whether it was suggested by someone else or was self-generated. In fact, for at least some levels of expertise, the source or explicitness of the diagnosis might be important in how thoroughly it is considered.

It is well documented that the probability rating assigned to a particular diagnosis tends to be greater when that diagnosis is presented in isolation relative to when it is presented within a list of alternative diagnoses (the unpacking effect).[2] For example, the rated probability that a person will die of cancer tends to be greater when cancer is considered by itself than when presented within a list of differential diagnoses. Our previous work has shown, somewhat counter-intuitively, that the alternative diagnoses that have the greatest influence on the probability assigned to a focal diagnosis are those that are most likely to have been considered even in the absence of their explicit presentation.[3] That is, the magnitude of the unpacking effect (i.e., the decrease in the probability assigned to the focal diagnosis) was greater when the unpacked alternatives were believed to be highly plausible by independent experts relative to when the unpacked alternatives were believed to be less likely. While this result suggests that participants under-appreciated diagnostic alternatives that they themselves generated relative to when the same alternatives were explicitly presented, the experimental design did not allow us to be certain that participants had actually considered the diagnoses that were rated as most likely. Five diagnoses were explicitly presented in the unpacked condition, thereby allowing the possibility that participants had not generated all of the plausible diagnoses while reading the case history.

The current study was designed to demonstrate the same result for alternatives that participants claimed to have actually considered. Furthermore, we attempted to maximize the probability that a specific alternative diagnosis would come to mind even when not presented explicitly by using clinical cases previously shown to have two highly likely and roughly equiprobable diagnoses.[4] Both manipulations should eliminate the unpacking effect if diagnosticians evaluate diagnostic possibilities that they themselves generate in the same way as diagnoses that are explicitly provided.

While subjective estimates of probability are believed to provide a valid measure of participants' clinical decision-making processes, it is possible that the act of assigning probabilities is a formal exercise that is not closely related to actual practice. So, the current study also served as an attempt to demonstrate that the unpacking effect is not restricted to numerical estimates of probability by examining whether or not patient management strategies, such as the requesting of diagnostic tests, are influenced by the explicit presentation of diagnostic alternatives. That is, if diagnosticians request more tests upon being presented two highly diagnostic alternatives relative to being presented just the focal diagnosis, we would have converging, and perhaps more ecologically valid, evidence that there is a tendency to under-weight alternatives that are not explicitly provided. Redelmeier et al.[5] have previously shown that the likelihood that fourth-year medical students will order a CT scan upon the presentation of a potential case of sinusitis was influenced by the number of alternative diagnoses that were explicitly mentioned. The current study attempts to further ensure the robustness and generalizability of their findings by using multiple cases and a more extreme manipulation.

We tested this design initially on medical students primarily because of the ease of obtaining them as participants, but we believe that this initial step provides data of interest. Numerous biasing studies in medicine have confirmed that both experts and novices tend to be susceptible to the same heuristic-induced errors.[1,5,11] Understanding the mechanism underlying such processes might allow insight into the source of any errors that are made even as the number of errors a diagnostician makes undoubtedly decreases with the development of expertise.

Method

Participants. The participant pool for this study consisted of second-year medical students from McMaster University's graduating class of 2001. A sample of tutorial leaders asked their students whether they would participate. Those who agreed were run through the experiment in their tutorial groups during two sessions separated in time by an average of eight days (range 4–14 days). Twenty students participated in four groups, but follow-up data could not be collected for one of the students, leaving 19 with complete sets of data. Upon completion of the second group session, participants were paid $20 and given feedback on both the clinical cases used and the purpose of the study.

Materials. Participants were presented ten case histories, each of which was followed by one or two diagnostic hypotheses and a series of five questions. (1) "Given the case history that you have just read, please assign a number between 0 and 100 indicating how likely you think it is that the case history is representative of the given diagnosis(es)." In all conditions participants were told that the diagnoses were mutually exclusive and that the inclusion of an "all other diagnoses" alternative meant that each list was exhaustive, thereby indicating that the sum of the ratings assigned should be 100%. (2) "Are there any diagnostic tests that you would like to see performed to aid you in your decision? If yes, please list them." (3) "While reading the case history, did you consider any diagnosis apart from those listed above? If yes, please state the diagnosis that you consider to be the most likely differential." (4) "Please rate your confidence (on a scale of 1 to 100) that you know the correct diagnosis." (5) "Please rate the typicality of this case on a scale of 1 to 100." The latter two questions were intended to serve as dummy variables that would increase the likelihood that participants would not remember the exact probability assigned to any particular question.

Procedure. Each of the ten cases has been shown to be suggestive of two diagnoses, both highly likely and roughly equiprobable

diagnoses.[5] One of each pair of diagnoses was randomly selected to be the focal diagnosis—the diagnostic alternative that would be presented with its associated case history across all conditions. In working through all ten cases, each subject was shown five cases within each condition (i.e., focal diagnosis alone versus focal + alternative diagnosis) randomly mixed together. Approximately one week later each participant was shown the same ten cases and asked to rate the original alternative(s) together with the alternative they had generated in response to question three. If no alternative had been generated, participants were simply shown the original alternatives a second time. Apart from adding the alternatives participants had generated during the first pass, the questionnaires used during the two sessions were identical.

Results

In completing all ten cases, 190 observations were generated that could be analyzed for the unpacking effect—a decrease in the probability assigned to a focal diagnosis upon the explicit presentation of additional diagnoses. Table 1 presents the average probability assigned to the focal diagnosis as a function of condition. First, a 2 (session) × 2 (number of alternatives presented during pass 1) × 2 (diagnostic alternative: generated versus not generated) × 10 (case) ANOVA was performed. A significant effect of "number of alternatives" ($F(1,156) = 12.365$, $p < .01$) revealed that the probability assigned to the focal diagnosis was higher when presented in isolation than it was when presented in conjunction with a second diagnosis even though the alternative diagnosis was the most likely differential. An effect of "diagnostic alternative" was also found ($F(1,156) = 6.009$, $p < .02$), thereby indicating that participants rated the focal diagnosis as more likely when they did not generate a plausible alternative diagnosis, as indicated by their responses to question 3. Case was the only other effect that reached significance ($F(9,156) = 4.746$, $p < .01$).

To further demonstrate that the unpacking effect occurs even when the unpacked alternatives are diagnoses that the participants had already considered, we performed a 2 (session) × 2 (number of alternatives presented during pass 1) × 10 (case) ANOVA on only those observations in which a diagnostic alternative had been generated, that is, using only the data presented in the Alternative Generated column of Table 1. The main effect of "number of alternatives" persisted ($F(1,139) = 8.861$, $p < .01$). In addition, a main effect of session was found ($F(1,139) = 16.375$, $p < .01$), which indicates that the probability assigned to the focal diagnosis was lower in session 2 than in session 1 even though the only difference between the two sessions was the explicit presentation during session 2 of the diagnoses that the participants claimed to have considered implicitly during session 1. Case was, once again, the only other effect that achieved significance ($F(9,139) = 4.323$, $p < .01$). The effect of session was not observed when the same analysis was repeated for trials in which the participants did not generate a diagnostic alternative (i.e., using only the data presented the Alternative Not Generated column of Table 1). This indicates that the effect was not simply a result of the passage of time. The numbers of observations in these cells were small, but examination of the means suggests that, if anything, the probability assigned to the focal diagnosis increased in session 2 relative to session 1 if no diagnostic alternative had been generated during session 1 ($F(1,17) = 0.017$, $p > .85$).

We also examined whether or not the phenomenon being illustrated by the probability ratings might influence management strategies by asking our participants to list the diagnostic tests that they would be interested in seeing performed. Participants requested more tests when two diagnoses were presented (mean = 3.464) than when the focal diagnosis was presented in isolation (mean = 2.989; $F(1,187) = 4.938$, $p < .05$). This result suggests that the explicit presentation of diagnoses can influence the management strategies of diagnosticians in addition to altering their rating of another diagnosis's likelihood.

Table 1. Mean Probability Ratings (and Counts of Numbers of Observations) Assigned to the Focal Diagnosis across Condition, by Second-year Medical Students at McMaster University, 1998–99*

	Focal Diagnosis		
Diagnosis(es) Presented	Alternative Generated	Alternative Not Generated	Overall
Session 1			
Focal	44.39 (89)	60.83 (6)	45.43 (95)
Focal + alternative	33.86 (70)	38.80 (25)	35.16 (95)
Overall	39.75 (159)	43.06 (31)	40.29 (190)
Session 2			
Focal (+ generated alternative if generated)	37.13 (89)	59.17 (6)	38.53 (95)
Focal + alternative (+ generated alternative if generated)	30.07 (70)	42.00 (25)	33.21 (95)
Overall	34.03 (159)	45.32 (31)	35.87 (190)

*Each of 19 medical students reviewed ten clinical cases in sessions one week apart. See text for details.

Finally, the effect of the number of alternatives presented on confidence ratings and typicality ratings were analyzed. No effect of session or "number of alternatives" was found for either of these two variables.

Discussion

These results support the notion that individuals tend to underappreciate self-generated diagnoses relative to diagnoses that are explicitly presented. The participants rated the originally presented diagnosis as less probable when the alternative they had claimed to be considering implicitly was provided in a more explicit manner. That is, the unpacking effect was found even when the diagnostic alternative that was unpacked was one that our participants claimed to have considered while originally viewing the case.

While differences in the probability ratings assigned to the focal diagnosis across condition might appear small relative to the 100-point scale used, it is important to note that the functional range of potential responses was actually substantially smaller than 100. As mentioned earlier, the cases were originally designed to be indicative of two diagnoses, which are both highly likely and roughly equiprobable. Consistent with that manipulation, our participants were hesitant to assign a very high or a very low likelihood rating to any individual diagnosis. The effect size across packed versus unpacked versions of the questionnaire was 0.46—a medium-sized effect[6]—even though substantial effort was invested to ensure that the cards were stacked in favor of the null hypothesis.

That being said, the mechanism that causes individuals to underweight alternatives that are not explicitly presented remains in question. As alluded to in the introduction, the effects observed might arise as a result of confirmation bias, as the explicit presentation of a diagnostic alternative might cause diagnosticians to differentially process the evidence relevant to the diagnostic possibilities. This could arise in at least two ways that are not necessarily exclusive of one another; the explicit presentation of a diagnostic hypothesis might influence both the search for and the construal of evidence. Support for the plausibility of these hypotheses is widespread.

For example, it has been found that, when given the opportunity to select additional information (i.e., prevalence data), medical students,[9] residents,[10] and physicians[11] tend to seek data that are relevant to a single disease while ignoring information that is related to equally plausible differential diagnoses. This biased search for

information need not be proactive in that it does not necessarily take place while the diagnostician gathers novel information. On the contrary, Anderson and Pichert have shown that retrieval of information from memory is also influenced by the context within which the search takes place.[12] When asked to recall information about a house, the type of information participants were able to remember was dependent on whether they had been asked to read the story from the perspective of a burglar or a home buyer. When subjects were later asked to adopt the opposite perspective, they were able to recall more information that simply had not been available during the first memory task. A plausible extension of this result is that the explicit presentation of a diagnosis might bias the memorial retrieval of features present in the case history.

Furthermore, maintaining an initial focus on the diagnosis that is explicitly presented might make it difficult to realize that nondiscriminating symptoms provide evidence for more than one diagnostic alternative. For example, considering the nausea and vomiting with which an 18-year-old woman with right-lower-quadrant discomfort presents as indicative of appendicitis might blind an individual to the possibility that these symptoms can also be construed of as clinical manifestations of pelvic inflammatory disease. Norman, LeBlanc, and Brooks have provided evidence that supports this notion by reporting that the mere presentation of a diagnostic alternative can influence the interpretation of classic clinical features.[7] Reinterpreting these features in light of self-generated diagnoses could prove to be difficult.

Regardless of their cause, the data presented here indicate that the meaning of the verb "to consider" should not necessarily be taken at face value. Having considered the plausibility of a diagnostic alternative can mean anything from having had the term come to mind to having performed a comprehensive analysis of the evidence for and against that particular diagnosis. Asking our participants to assign a probability rating to the likelihood of diagnoses that they claim to have considered was sufficient to decrease the probability that they were willing to assign to the focal diagnosis. This strongly suggests that the evidence in favor of the self-generated alternative was underappreciated relative to when attention was focused on that alternative explicitly. Further research is required to determine whether or not particular strategies can be adopted to prevent such under-weighting.

The authors thank John Cunnington, Rose Hatala, Alan Neville, Geoff Norman, Richard Reznick, and anonymous reviewers for useful comments and discussion during the development and completion of this project.

Correspondence: Kevin Eva, Department of Psychology, McMaster University, Hamilton, Ontario L8S 4K1, Canada; e-mail: (evakw@mcmaster.ca).

References

1. Reisberg D. Cognition: Exploring the Science of the Mind. New York: W. W. Norton & Company, 1997.
2. Tversky A, Koehler DJ. Support theory: a nonextensional representation of subjective probability. Psychol Rev. 1994;101:547–67.
3. Eva KW, Brooks LR, Cunnington JPW, Norman GR. The strength of alternatives: its role in diagnostic decision making and probability judgments. Unpublished paper.
4. Cunnington JPW, Turnbull JM, Regehr G, Marriott M, Norman GR. The effect of presentation order in clinical decision making. Acad Med. 1997;72(10 suppl 1):S40–S42.
5. Redelmeier DA, Koehler DJ, Liberman V, Tversky A. Probability judgment in medicine: discounting unspecified probabilities. Med Decis Making. 1995;15:227–31.
6. Cohen J. Statistical Power Analysis for the Social Sciences. 2nd ed. New York: Academic Press, 1977.
7. Norman GR, LeBlanc VR, Brooks LR. On the difficulty of noticing the obvious. Psychol Sci. In press.
8. Hatala R, Norman GR, Brooks LR. The impact of a clinical scenario upon accuracy of electrocardiogram interpretation. Unpublished paper.
9. Kern L, Doherty ME. 'Pseudodiagnosticity' in an idealized medical problem-solving environment. J Med Educ. 1982;57:100–4.
10. Wolf FM, Gruppen LD, Billi JE. Differential diagnosis and the competing-hypotheses heuristic: a practical approach to judgment under uncertainty and Bayesian probability. JAMA. 1985;253:2858–62.
11. Green LA, Yates JF. Influence of pseudodiagnostic information on the evaluation of ischemic heart disease. Ann Emerg Med. 1995;25:451–7.
12. Anderson RC, Pichert J. Recall of previously unrecallable information following a shift in perspective. J Verbal Learn Behav. 1978;17:1–12.

94

The Impact of Structured Student Debates on Critical Thinking and Informatics Skills of Second-year Medical Students

STEVEN A. LIEBERMAN, JULIE M. TRUMBLE, and EDWARD R. SMITH

Yet it has become increasingly difficult to keep abreast of and to assimilate the investigative reports which accumulate day after day. . . . (O)ne suffocates . . . through exposure to the massive body of rapidly growing information.

—BERNHARD VON LANGENBECK, Address at the First Congress of Surgery, April 10, 1872[1]

Among its many facets, the field of medical informatics encompasses the use of technology to access and manage scientific information. The Association of American Medical Colleges (AAMC), through the Medical Informatics Objectives of the Medical School Objectives Project (MSOP) has identified five informatics-related roles of the physician and has established objectives for each of these roles. The lifelong learning role incorporates skills relating to information retrieval, evaluation, and reconciliation. Without activities specifically targeting these skills, it is an act of faith that students will graduate with adequate preparation in these areas.

To explicitly address these curricular goals, second-year students in Endocrinology and Reproduction Course at the University of Texas Medical Branch in Galveston were required to participate in debates on controversial topics in these fields. This exercise provided a structured task for developing and improving skills in literature searching, critical thinking, including evaluation of the quality of studies, reconciling results of conflicting studies, teamwork, formal presentation and communication, and spontaneous scholarly discussion.

A search of the Medline database produced only one article describing the use of student debates for acquiring content and developing critical thinking and communication skills in health science education.[2] The paper describes a first-year pharmacy curriculum that incorporated debates on socioeconomic topics relevant to pharmacy practice. While these debates required critical analysis of issues, the primary focus was on content rather than cognitive or informatics-related skills.[2]

Published accounts of debates in a college chemistry course[3] and business school[4] provide qualitative descriptions of the beneficial effects of such exercises on critical thinking, updating knowledge, and communication skills. In a more quantitative approach, Allen et al.[5] conducted a meta-analysis of the impact of formal instruction in communication skills (including debates) on critical thinking ability. Such training resulted in 44% increase in scores on tests of critical thinking. Compared with training in other communication skills, participation in "forensics" (i.e., competitive debates) produced the greatest improvement, although the differences did not achieve statistical significance.[5] Finally, Johnson et al.[6] performed a meta-analysis of the effects of a method they call "academic controversy" on a variety of cognitive outcomes. This method, which shares many features of the debates described in the current report, has produced "increased achievement and retention, higher-quality problem-solving and decision-making, more frequent creative insight, more thorough exchange of expertise, and greater task involvement" by students.[6]

The current report describes the implementation of structured debates and the evaluation by students and faculty of the degree to which the informatics objectives were accomplished.

Method

The 174 second-year students were divided into six sections of approximately 30 students each and were assigned to teams of three within each section. The debate topics represented areas of controversy in endocrinology and reproductive science. Students received assigned topics at the beginning of the course, and each student participated in one debate. When not presenting, students were expected to attend their section's debates. Each team researched background information, identified the main issues, found and analyzed relevant studies, developed arguments on both sides of the topic, developed a "rational compromise" after weighing the evidence, and prepared to present each side of the topic.

Each team was assigned to present the pro (supporting) or con (opposing) perspective immediately before the debate. Each of the six students gave a five-minute presentation of one of the following segments: pro background and arguments; pro supporting data; con background and arguments; con supporting data; pro "rational compromise"; con "rational compromise." All team members participated in a ten-minute rebuttal segment prior to the "rational compromises" and a ten-minute question-and-answer session after the final presentation. A faculty moderator kept the session on schedule, participated in the question-and-answer portion, and evaluated the students' performances.

The students were assessed individually on presentation skills, contributions to the rebuttal and question-and-answer segments, and professionalism. The teams were evaluated for the accuracy of information and appropriateness of conclusions, and on written summaries and references turned in at the debate. Individual and team scores were combined to generate a letter grade for each student.

Debates presented in each of the six student sections in a given week generally revolved around a single theme in order to provide similar learning experiences for all students attending. For example, one set of debates addressed related facets of the role of insulin resistance in producing disease: (1) Hyperinsulinemia causes hypertension; (2) Insulin resistance increases the risk of thrombosis; (3) Insulin resistance increases the risk of coronary artery disease; (4) Insulin resistance causes the polycystic ovary syndrome; (5) Obesity s an independent risk factor for coronary artery disease; (6) Intensive treatment of type 2 diabetes mellitus lowers the risk of coronary artery disease compared with conventional treatment. Other themes included menopausal hormone replacement therapy, HIV infection and pregnancy, growth hormone therapy in non-growth-hormone-deficient children, and the diagnosis and management of thyroid and parathyroid neoplasia.

The effectiveness of this exercise was evaluated by three modalities. First, all students (n = 174) were requested to complete a survey following their debates. Second, faculty moderators (n = 17) were surveyed to obtain their impressions of the students' skills and the educational value of the debates. Finally, two focus groups of randomly selected students (n = 4 per group) met with facilitators midway through the course to discuss the debates and other course aspects. The facilitators were educators not directly involved in the course. Summaries and anonymous comments from the focus groups were reviewed and approved by the students.

The student and faculty questionnaires were parallel instruments

95

TABLE 1. Student Self-assessments and Faculty Ratings of Skills Developed during Preparation and Presentation of Structured Student Debates, University of Texas Medical Branch at Galveston, 1999–2000*

Skill	Student Self-assessment				Faculty Rating
			Percentage of Students Improving by:		
	Before	After	1 Level	≥2 Levels	
A. Weighing conflicting information from multiple sources and reconciling the differences (MSOP Informatics Objective A.3.c)	3.23 (0.13)	4.31 (0.10)†	32.5%	30.8%	3.80 (0.49)
B. Critically reviewing published research (MSOP Informatics Objective A.3.d)	3.09 (0.12)	3.97 (0.08)†	34.2%	23.9%	3.00 (0.45)
C. Discriminating between types of information sources in terms of currency, format, authority, relevance, and availability (MSOP Informatics Objective A.3.b)	3.23 (0.12)	4.15 (0.08)†	29.9%	26.5%	3.80 (0.37)
D. Recognizing factors that influence the accuracy/validity of information (MSOP Informatics Objective A.3.a)	3.26 (0.12)	4.17 (0.08)†	29.1%	25.6%	3.40 (0.40)
E. Making evidence-based decisions (MSOP Informatics Objective A.4.c)	3.62 (0.12)	4.39 (0.08)†	30.8%	20.5%	3.20 (0.49)
F. Expressing the relative risks and benefits of outcomes/treatment options (MSOP Informatics Objective B.5.b)	3.41 (0.12)	4.22 (0.11)†	28.2%	22.2%	3.83 (0.17)
G. Medline searching (MSOP Informatics Objective A.2.a & b)	3.72 (0.13)	4.65 (0.08)†	18.8%	28.2%	4.50 (0.22)
H. Knowledge of cost–benefit issues in health care (MSOP Informatics Objective E.2.a)	2.94 (0.12)	3.67 (0.11)†	28.2%	18.8%	3.33 (0.33)
I. Ability to make formal presentations (MSOP Informatics Objective C.2)	3.75 (0.12)	4.35 (0.08)†	29.9%	14.5%	5.00 (0.00)
J. Maintaining a healthy skepticism about the quality of information (MSOP Informatics Objective A.4.b)	3.73 (0.12)	4.42 (0.09)†	20.5%	20.5%	3.17 (0.54)
K. Using multiple sources for problem solving (MSOP Informatics Objective A.4.a)	3.93 (0.11)	4.49 (0.08)†	28.2%	12.8%	4.17 (0.31)
L. Impromptu reasoning skills (the ability to "think on your feet")	3.82 (0.11)	4.15 (0.09)†	23.9%	5.1%	4.17 (0.31)
M. Working effectively as a team to accomplish tasks	4.57 (0.09)	4.91 (0.08)†	23.9%	4.3%	5.33 (0.21)

*Students were asked to indicate on a scale of 1 through 6 (1 = complete novice; 2 = minimally competent; 4 = moderately competent; 6 = expert) their "skill level on each of the following both BEFORE and AFTER the debate." 114 of 174 students (65.5%) completed the survey. Data are means ± SEM.

†$p < 0.0001$ by paired t test for comparisons of "before" versus "after."

and were administered following the debates. Section A asked the students to use a three-item scale—"major resource," "minor resource," and "not used"—to describe the importances of ten resource types. Faculty had one additional category, "can't judge." Section B addressed 13 specific objectives of the debates, 11 of which corresponded to skills identified in the MSOP (Table 1). Students used a scale from 0 to 6 (0 = not used/not applicable; 1 = complete novice; 2 = minimally competent; 4 = moderately competent; 6 = expert) to retrospectively rate their pre- and post-debate skills. The faculty scale replaced "not used" with "can't judge." Section C used Likert-like scales to assess the importances of skills fostered by the debates, and the usefulness and timing of debates for promoting skill development. Finally, section D asked for comments and suggestions.

Results

Of the 174 participants, 114 (65.5%) responded to the survey. They did not differ from the non-respondents with regard to individual debate scores (33.6 ± 0.2 versus 33.6 ± 0.3, $p > .9$), team debate scores (29.2 ± 0.3 versus 29.0 ± 0.4, $p > .7$), or scores on the final course exam (90.3 ± 0.8 versus 88.3 ± 1.4, $p > .2$). Six faculty, three clinicians and three basic scientists, who had moderated 19 of the 30 debates (63.3%), responded to the faculty survey. These six included all four who had moderated more than one debate.

Among the students responding, 78 (67%) indicated that the skills acquired through the debates would be "important" or "very important" in their careers, while all six faculty rated the importance of these skills in the highest category. Seven students (6%) felt the skills would be "not important at all." Seventy students (60%) agreed or strongly agreed that the debate had been a valuable learning exercise, while 23 (20%) disagreed or strongly disagreed. The students were evenly divided as to whether one (n = 33), two (n = 35), or three-to-four (n = 33) similar exercises would be required to "promote adequate development" of the skills. Four faculty (66.7%) felt three or four times would be appropriate, one felt four to eight times would be needed, and one felt two times would be adequate. One faculty member and 61 students (52%) felt the preclinical years were the most appropriate place in the curriculum for such exercises. Seventeen students (15%) felt they should be limited to the clinical years, and five faculty (83%) and 23 students (20%) indicated that the exercises should occur throughout the four-year curriculum.

The results of the student and faculty surveys of skill development are presented in Table 1. The students' self-assessments increased significantly for all skills, with mean ratings of post-debate skills generally near a score of 4, or "moderately competent." However, the increase in mean score was greater than one level for only one skill (weighing and reconciling conflicting information), while for two skills (impromptu reasoning; working effectively as a team) less than 40% of the respondents reported any increase. Although the sample sizes (114 students, 6 faculty) preclude statistical comparisons, faculty ratings of student skills appeared lower than student self-ratings for all but four skills.

Table 2 summarizes the students' responses regarding resource utilization. Review articles (88.9%) and primary research articles (86.3%) were most frequently identified as major resources.

Focus-group summaries corroborated the generally favorable survey findings. Specifically, the debates were perceived more as exercises in critical thinking than as exercises in content acquisition, had been effective in promoting literature-searching and research-analysis skills, and had been "interesting and enjoyable." The most common criticism was the amount of preparation time required. Comments from the faculty survey, while overall extremely favorable, suggested several areas for improvement: students' overreliance on reviews and published expert opinion, a tendency for students to want "to win" the debate rather than come to a balanced judgment based on the evidence, and the need to couple specific instruction in these skills with the debates.

Table 2. Numbers and Percentages of Second-year Medical Students Rating Information Resources as Major or Minor in Preparing for Structured Debates, University of Texas Medical Branch in Galveston, 1999–2000*

	Ratings					
	Major Resource		Minor Resource		Not Used	
Resource	No.	%	No.	%	No.	%
Review articles	104	88.9	12	10.3	1	0.9
Primary research articles	101	86.3	15	12.8	1	0.9
Systematic reviews/meta-analyses	52	44.4	46	39.3	19	16.2
Practice guidelines/consensus statements	28	23.9	48	41.0	41	35.0
Other textbooks	23	19.7	66	56.4	28	23.9
Required course textbook	15	12.8	80	68.4	22	18.8
Professional Internet sites	13	11.1	40	34.2	64	54.7
Consultation with an expert	6	5.1	53	45.3	58	49.6
Governments Internet sites	5	4.3	21	17.9	91	77.8
Commercial Internet sites	4	3.4	34	29.1	79	67.5

*Students were asked "Please indicate the importance of each of the following resources in preparing for your debate." Of 174 students participating in the debates, 114 (65.5%) responded.

Discussion

This report describes the method and evaluation of struct ed student debates for promoting the development of several cognitive and informatics-related skills, many of which are embodied in the MSOP. The data reported confirm that this exercise accomplished most of its goals. The central goal of promoting the development of skills in analyzing research studies and weighing and reconciling contrasting results was realized. The specific objectives related to this goal (A–D in Table 1) showed the greatest mean increases in self-ratings as well as the greatest proportions of students reporting improvement. Furthermore, primary research articles were among the two most important resource categories, corroborating the value of this exercise in stimulating critical analysis of research reports. However, the comparable emphasis on review articles raises concern that the exercise could deteriorate into general summaries rather than critical evaluations of the literature. In order to focus students' attention on the primary literature, the debate format and evaluation emphasized the use of data to support arguments. Some reliance on review articles was to be expected, as most students had neither extensive backgrounds in the topics nor much experience in reconciling conflicting research. Faculty impressions were lower than the students' self-ratings in these areas, especially with regard to "critically reviewing published research," with mean ratings below the "moderately competent" level. Thus, at the completion of this exercise, the faculty perceived lower abilities and, therefore, a greater need for further skill development than did the students.

The students also indicated improvement in literature searching, weighing risks and benefits of treatments, making evidence-based decisions, and understanding cost–benefit issues. For the other self-rated skills there were lower proportions of students improving and smaller increases in mean scores, although all increases were statistically significant. Although faculty assessments of most skills were lower than students', faculty rated the students at comparable or higher levels in literature searching, presentation skills, impromptu reasoning, and teamwork.

The retrospective nature of the student survey, in which the students rated both their pre- and post-debate skills after completing the debate, may be viewed as a weakness in the study design. Nonetheless, the increase in scores indicates that the students felt the exercise did, in fact, promote skill development. While the significant increases in mean scores indicate progress in students' skill development, the magnitudes of changes were generally small and the percentages of students reporting improvement varied by skill. These findings suggest that one such exercise is insufficient for adequate skill development. All faculty and most students acknowledged the importance of these skills and indicated that additional exercises were necessary. Consensus among faculty was for three or four exercises throughout the four-year curriculum, while the students' varied recommendations are best summarized as two debates during the preclinical years.

In summary, we have found that structured student debates among second-year medical students promoted development of critical thinking and informatics skills identified in the MSOP Medical Informatics Objectives. A series of exercises distributed throughout the curriculum, targeting progressively more advanced skills and coupled to instruction in these skills, may achieve these objectives more fully.

Correspondence and requests for reprints: Steven A. Lieberman, MD, Department of Internal Medicine, University of Texas Medical Branch, 301 University Blvd., MRB 8.138, Galveston, TX 77555-1060.

References

1. Strauss MB. Familiar Quotations. Boston: Little, Brown and Company, 1968.
2. Poirier S. Active involvement of students in the learning process of the American healthcare system. Am J Pharmaceutical Educ. 1997;61:91–7.
3. Streitb ger HE. A method for teaching science, technology, and societal issues in introductory high school and college chemistry classes. J Chem Educ. 1988;65:60–1.
4. Schroeder H, Ebert DG. Debates as a business and society teaching technique. J Business Educ. 1983;58:266–9.
5. Allen M, Berkowitz S, Hunt S, Louden A. A meta-analysis of the impact of forensics and communication education on critical thinking. Communication Educ. 1999;48:18–30.
6. Johnson DW, Johnson RT, Smith KA. Academic Controversy: Enriching College Instruction Through Intellectual Conflict. ASHE–ERIC Higher Education Report. 3rd ed. Washington, DC: The George Washington University Graduate School of Education and Human Development, 1996:123.

97

Critical Appraisal Turkey Shoot: Linking Critical Appraisal to Clinical Decision Making

ALAN J. NEVILLE, HAROLD I. REITER, KEVIN W. EVA, and GEOFFREY R. NORMAN

Since the publication of Physicians for the Twenty-First Century —"the GPEP Report" of 1984, medical educators have identified the need for physicians to become lifelong learners.[1] Part of the impetus for this conclusion arises from several studies that have demonstrated that knowledge and/or competence of physicians decline as a function of time since graduation; the evidence indicates the cause to be failure to acquire new knowledge rather than a tendency to forget previously learned material.[2] Thus, physicians need to be trained to identify the relevant medical literature (i.e., information-seeking skills) and to apply "critical appraisal" techniques to analyze potentially useful articles culled from the literature search.

There is little published evidence that educational interventions around critical appraisal teaching in undergraduate or postgraduate medical curricula impact in a sustained way the knowledge of epidemiologic principles or the critical application of current research information for clinical decision making.[3] In considering the impact on conceptual knowledge, one could argue that there is a lack of validated tools available for evaluating critical appraisal skills; alternatively, the format of instruction, timing in the curriculum, and duration of instruction may be at fault. More important, studies have not addressed the issue of whether the demonstration of mastery of particular critical appraisal skills can be related to clinical decision making. Ultimately, such mastery becomes largely irrelevant if it does not translate into better judgment.

The authors of this study were concerned that, despite the inclusion in the first-year undergraduate curriculum of several focused objectives surrounding critical appraisal in the domain of clinical epidemiology, feedback from clinical faculty suggested that students had only rudimentary knowledge of the application of these principles at the end of the first year. In contrast to this feedback, problem-based learning (PBL) is believed to hold the potential to equip graduates with the skills to learn after graduation. In fact, several studies have shown significant differences between students of PBL and students of conventional curricula in the use of recently published medical literature.[4–6] With this inconsistency in mind, two experimental questions were asked.

1. Are critical appraisal concepts to which students are "exposed" in PBL in earlier curricular blocks retained sufficiently to allow identification of methodologic errors in formal articles?
2. Does awareness of such methodologic flaws transfer to an appreciation of how these errors might invalidate the conclusions of the journal articles' authors?

Ergo, the goal of this study was to investigate the relationship between understanding the concepts of critical appraisal and their application in clinical decision making. Understanding this relationship can potentially improve the teaching of critical appraisal and the evaluation of this teaching.

Methods

Participants. This was a single-blinded experimental design study. The participant pool was composed of two consecutive first-year undergraduate medical school classes (the graduating classes of 1999 and 2000, respectively) in a PBL curriculum at McMaster University. Each class was composed of 18 tutorial groups of five to six students each. The students had some background in critical appraisal, as it had been studied in a readily identifiable manner during the first curricular unit at the beginning of the first academic year. For each class, the study took place during the third curricular unit running during the final three months of their first academic year.

Materials. The subunit planners for each month-long subunit in that third curricular unit selected two journal articles from their respective expert domains of gastroenterology, hematology, and endocrinology. These context experts chose articles that met the defined criteria of being (a) methodologically sound and (b) not directly covered within the context of the unit's curricular problems. Within each of the six articles so identified, one, two, or three different methodologic flaws were implanted, each flaw sufficiently egregious to warrant dismissal of the author's conclusions. The methodologic flaws inserted related to concepts that students were expected to have come across previously in the curriculum. Six categories of errors were examined (participant assembly, randomization, contrast, follow-up, analysis, and other). For example, the study group may have been inappropriately pooled or randomization might have been non-blinded. The text of the journal articles was retyped with the titles, tables, authors, and journal names absent. After this was done, the original six "gold" articles and their mirror flawed counterparts, or "turkey" articles, were superficially indistinguishable.

For each of the six articles a related clinical scenario was generated that would present a clinical management problem for which a specific intervention was to be considered. Each problem was relevant to the unit of study but was not directly related to the health care problems in the curriculum and could not be answered using standard textbooks. Also, according to the subunit planners, the answers to the problems should have been obvious if the relevant recent literature was known.

Procedure. Within both the class of 1999 and the class of 2000, students were randomly allocated biweekly to receive either a gold or a turkey article, for a total of six articles over 12 weeks. Randomization took place across the entire class, not by tutorial group, since the students worked on the exercise independently, and assignment was by use of a table of random numbers. The students were all given a "pre-appraisal" response sheet with the appropriate clinical scenario and were asked to respond on an anchored seven-point Likert-type scale whether they agreed or disagreed with the optional management or intervention suggested. The scale was anchored between "definitely yes" (1), "probably yes" (2), "probably no" (5) and "definitely no" (7). This pre-appraisal response sheet served as a baseline of the students' knowledge of the condition demonstrated by the scenario. The students were then given two weeks to work on the articles they had been assigned. Afterward, the students completed a "post-appraisal" sheet that presented the same clinical scenario and the same clinical question that they had seen two weeks earlier. In addition, they were asked to identify any methodologic flaws in the articles they had read. For the class of 1999, this identification took place using an open format. For the class of 2000, the identification of flaws was noted by ticking them off a checklist that contained 29 potential methodologic errors, three to six per category. Responses to the post-appraisal questionnaire would allow us to estimate the students' ability to detect methodologic flaws and to assess whether or not the author's con-

clusions had influenced their clinical decisions. The responses were handed in to the tutor and a "tutor-guide" was provided to briefly explain the inserted flaws, thereby allowing discussion of the critical appraisal issues during tutorials.

Results

Eighty-nine of the 100 students in the class of 1999 completed both the pre- and the post-appraisal questionnaires for at least one of the six questions. The average number of completed questions per participant was 5.61, with 69 of the 89 students completing all six questions. In the class of 2000, 63 of the 100 students completed both pre- and post-appraisal questionnaires at least once, averaging 5.68 questions per participant, with 50 of the 63 students completing all six questions. The decreased participation by students in the second year reflected ambivalence on the part of some of the tutors in dealing with the logistics of the exercise. Two hundred and forty-six (49.3%) of the 499 observations collected from the class of 1999 and 186 (52.0%) of the 358 observations collected from the class of 2000 were from the gold arm of the studies, thereby indicating that the questions were not completed differentially for the two types of papers provided.

Table 1 presents the mean pre-test and post-test scores for both the turkey and the gold groups of both classes. Upon coding the data, some scales were reversed so that the low end of the seven-point Likert scale was always the "correct" response. In neither class did the pre-test scores of the two groups differ significantly from one another. A 2(time: pre- vs. post-) × 2(arm: gold vs. turkey) repeated-measures analysis of variance revealed a significant interaction between time and arm ($F(1,497) = 7.043$, $p < (.01)$ for the class of 1999. The same analysis revealed an effect that bordered on significance for the class of 2000 ($F(1, 356) = 3.273$, $p < .075$). Planned comparison t-tests for both classes revealed the nature of these interactions. Mean post-test scores of both gold groups decreased significantly relative to their pre-scores ($t[245] = 5.198$, $p < .01$ and $t[185] = 4.834$, $p < .01$ for the class of 1999 and the class of 2000, respectively). In contrast, mean post-test scores of both turkey groups did not reveal a significant effect of time ($t[252] = 1.323$, $p > .18$ and $t[171] = 1.693$, $p > .09$ for the class of 1999 and the class of 2000, respectively). Therefore, students were more likely to change their management decisions in an appropriate direction if they had read a methodologically error-free version of the paper.

The participants who read the error-free gold version of the article did report having found errors, as can also be observed in Table 1, but they reported having found significantly fewer errors than those who read the turkey version of the article ($t[496] = -3.252$, $p < .01$ and $t[357] = -3.338$, $p < .01$, for the class of 1999 and the class of 2000, respectively). Collapsing across arms, there was a significant positive relationship in both classes between the number of problems raised and the post-score assigned ($r = 0.230$, $p < .01$ for the class of 1999, $r = 0.344$, $p < .01$ for the class of 2000). This indicates that the fewer errors raised, the lower (i.e., more correct) the post-score that was assigned. This relationship remained significant when the analysis was limited to the correct identification of the errors that had been planted within the turkey articles ($r = 0.163$, $p < .01$ and $r = 0.251$, $p < .01$ for the classes of 1999 and 2000, respectively). These analyses provide converging evidence that students were altering their management decisions based on the strength of the method that they perceived. In addition, it is reassuring that the participants did not appear to allow their prior impressions of the appropriate management decisions to influence their critical appraisals of the articles presented. This is evidenced by the lack of a relationship between the number of problems raised and the pre-score assigned ($r = 0.016$, $p > .72$ and $r = 0.068$, $p > .19$ for the classes of 1999 and 2000, respectively).

Finally, taking into account the numbers of turkey articles read and the numbers of errors embedded, the potential numbers of errors that could be correctly identified were 505 and 343 for the classes of 1999 and 2000, respectively; 178 (35.2%) of them were identified by the class of 1999 and 80 (23.2%) by the class of 2000. Review of the actual methodologic flaws identified by the students demonstrated no consistent pattern between the two classes. The proportions of the six individual error categories correctly identified by the class of 1999 were 33/86 (38%) for participant assembly, 37/98 (38%) for randomization, 63/163 (39%) for contrast, 11/45 (24%) for follow up, 8/68 (12%) for analysis, and 26/45 (58%) for other. The corresponding proportions correctly identified by the class of 2000 were 13/56 (23%), 21/56 (38%), 26/113 (23%), 16/29 (55%), 4/59 (7%), and 0/30 (0%) for the same six categories, respectively.

Table 1. Mean Responses to Patient Management Problems by Class and Type of Article* McMaster University 1997 and 1998

Class	Arm	Pre-test Score	Post-test Score	No. of Errors Identified
1999	Gold	3.764	3.195	2.398
	Turkey	3.644	3.506	2.805
2000	Gold	3.460	2.929	2.355
	Turkey	3.496	3.293	3.255

*Study conducted on two consecutive classes of first-year students. For each of the classes, mean pre-test (before reading the articles) and post-test scores (seven-point scale) reflecting agreement with the articles' conclusions are given. Gold arm = students allocated the original articles; turkey arm = students allocated articles with methodologic flaws inserted.

Discussion

An ultimate objective in teaching critical appraisal concepts is for medical students to view literature searching and critical appraisal as fundamental skills required for effective medical practice. As Norman et al. demonstrated in a recent review of teaching critical appraisal, most reported teaching interventions, even the few controlled studies published, have assessed short-term gains in acquiring knowledge of critical appraisal techniques rather than their application to clinical decision making.[6] The results of these studies were largely consistent with the anecdotal feedback that we have received from tutors—students appear to be poor critical appraisers. While it is important to be able to demonstrate some knowledge of the principles of how to scrutinize the medical literature carefully and critically, some demonstration of putting these principles into practice would seem to be just as desirable an educational outcome. By using a more decision-oriented outcome measure, the current findings suggest that the studies reviewed by Norman et al. and the interactions between students and tutors might underestimate students' ability to critically appraise scientific articles.

This study demonstrated that first-year medical students can alter their clinical management decisions appropriately as a function of whether they have read a methodologically sound or flawed journal article. When provided with the "gold" journal articles, these students changed their clinical decisions in the post-test in the direction of the correct management decisions, despite apparently identifying some putative methodologic flaws in these "gold" papers. As expected, however, fewer errors were identified by students in the "gold" articles, and there was a significant positive relationship between identifying fewer errors and assigning a "more correct" clinical decision score on the post-test.

The findings from the turkey articles require more explanation. As expected, students identified more errors in the turkey papers. However, at most, only 35% of the deliberately inserted methodologic flaws were correctly identified. Despite being unable to ac-

99

curately identify all of these errors, the students tended not to alter their original management decisions when they had been assigned turkey papers. It seems that the students were uncomfortable with the authors' conclusions and, without necessarily being able to specify the flaws, decided to either maintain their original management decisions or make small changes in either direction. While the authors had anticipated from a curriculum review that the "flaws" inserted into the articles might be identified by students, one weakness of this study is that there was no assessment of the tutors' abilities to identify them.

Finally, there was no relationship between the number of flaws identified and the "correctness" of the scores the students assigned on the pre-test. This implies that the students were able to read the articles critically without being biased by their perceptions of the correct management decisions, thereby providing further evidence that our students treated the articles in a rational manner.

In summary, the current findings show that our first-year students do indeed have relatively limited ability to identify specific methodologic issues in journal articles. Despite this, however, the clinical decision-making results demonstrated a gratifying relationship between the students' perceptions of the "quality of evidence" and appropriate changes in their management decisions. This suggests that students are reading the literature more critically than might be assumed by simply testing their knowledge of particular critical appraisal concepts. That is, while seeming to treat articles appropriately, students may not be able to articulate specific methodologic errors, thereby giving the appearance of poor critical appraisal skills. While it is important for students to be able to articulate critical appraisal concepts, the current results suggest that examining students' abilities in this domain should take place in the context of clinical decision making. Our participants' capacity to alter their decisions in a rational manner suggests that even novice medical students should be strongly encouraged to critically appraise. Future research will determine to what extent the correct or incorrect perceptions by students of particular methodologic flaws influences their clinical decision making.

The authors thank Glenn Jones for generating the checklist that was used by the class of 2000 and Annette Schrapp for administrative support in preparing and distributing the materials to study participants and tutors.

Correspondence: Kevin Eva, Department of Psychology, McMaster University Faculty of Medicine, Hamilton, Ontario L8S 4K1, Canada.

References

1. Muller S (chairman). Physicians for the twenty-first century: report of the project panel on the general professional education of the physician and college preparation for medicine. J Med Educ. 1984;59(11 Pt 2).
2. Day SC, Norcini JJ, Webster GD, Viner ED, Chirico AM. The effect of change in medical knowledge on examination performance at the time of re-certification. Proc Annu Conf Res Med Educ. 1988;22:139–44.
3. Norman GR, et al. Effectiveness of instruction in critical appraisal (evidence-based medicine) skills: a critical appraisal. Can Med Assoc J. 1998;158:177–81.
4. Blumberg P, Michael J. Development of self-directed learning behaviours in a partially teacher-directed problem-based learning curriculum. Teach Learn Med. 1992; 4:3–8.
5. Marshall JG, Fitzgerald D, Busby L, et al. A study of library use in problem-based and traditional medical curricula. Bull Med Libr Assoc. 1992;81:299–305.
6. Shin JH, Haynes RB, Johnston ME. Effect of problem-based, self-directed undergraduate education on life-long learning. Can Med Assoc J. 1993;148:969–76.

100

Communication Skills in Medical School: Exposure, Confidence, and Performance

DAVID M. KAUFMAN, TONI A. LAIDLAW, and HEATHER MACLEOD

Numerous studies indicate that, although communication skills can be learned, they can also deteriorate as students progress through medical school, particularly in the clinical years as students learn medical problem solving.[1–5] The good news is that this deterioration in communication skills can be prevented or reduced with more rigorous training. This was the surprise finding of Davis and Nicholaou,[6] who compared the communication skills of first- and fourth-year medical students. They found that fourth-year students had superior facility in communication skills, which is attributable to a greater emphasis on the importance of communication and increased training in the curriculum. To be effective, communication training must provide bridges between theory, knowledge, practice, and exposure—with exposure providing students contact with patients through clinical observation and clinical consultation. Students acquire the most effective interviewing skills when they interact with patients during their clinical training,[7] so exposure to a wide variety of clinical situations is essential. Prior training for such clinical encounters helps students develop working knowledge, understanding, and communication skills for dealing with challenging doctor–patient interactions.[8] Students must fulfill three conditions to demonstrate appropriate communication skills.[7] First, they need to know and understand a minimum of the corpus of knowledge and theory underlying communication exchanges in general and consultation processes. Second, they need to have a positive attitude towards using these skills in their interactions with patients. According to Bandura,[9] this attitude is best developed through positive role modeling. Third, students need to be trained in a repertoire of specific communication skills and techniques and be placed in situations where these can be practiced successfully with patients.[10]

The purpose of this study was to examine students' exposures to and confidence in communication skills, the relationship between exposure and confidence, and the relationship between exposure and performance of patient–doctor communication skills among students graduating from an undergraduate medical program. By exposure we mean observing, assisting, or performing the skill. The four categories of communication skills we studied were interviewing, breaking bad news, crisis management, and counseling. We refer to the last three of these as "higher-order" skills, as they involve progressively more challenging and complex communication interactions.

Background

Preclerkship Curriculum (Years One and Two). A problem-based learning (PBL) curriculum was begun at Dalhousie University Faculty of Medicine in 1992. The primary vehicle used to instruct students in communication skills is a module on interviewing skills in the first-year Patient–Doctor unit. Students are videotaped interviewing a standardized patient, and they practice their skills in small groups. They also receive lectures and written material. The students observe and practice basic history taking in clinical settings in their first and second years.

Clerkship Curriculum (Years Three and Four). At the time of the study, the clerkship comprised an 86-week continuum of experience, with significant flexibility and student choice. The students received some formal training in communication skills during their family medicine and psychiatry rotations. However, in other rotations, instruction occurs in clinical settings on an ad hoc basis, without a formal curriculum, as needs are identified.

Method

The students in the sample comprised the first two classes (n = 172) to graduate from the new PBL curriculum at Dalhousie (classes of 1996 and 1997).

A locally-developed questionnaire was used to obtain students' self-assessments of exposure and confidence. It consisted of four sections that asked students to indicate their levels of exposure to a set of ten communication skills (see Table 1). They also were asked to rate their confidence with respect to each skill, using a 6-cm visual analog scale with the ends marked "low" and "high." This is a useful and rarely used approach to assessing students' confidence in their skills. The rating scale for exposure consisted of the categories: never encountered, observed only, assisted senior staff member, performed once, performed two or more times. Students in the classes of 1996 and 1997 also participated in a two-hour objective structured clinical examination (OSCE) with simulated patients. However, three ten-minute communication stations were added to the 1997 OSCE, dealing with (1) requesting an organ donation from the husband of a woman declared "brain dead," (2) counseling a middle-aged woman with depression, and (3) managing a 70-year-old woman brought to the emergency department by her daughter after a fall. All students were rated in each station by a trained physician–examiner, using a standard rating scale.

The students took the two-hour OSCE on the day following completion of their final clerkship rotations at the end of medical school. While awaiting their results in a large room, they completed a series of questionnaires, including the one used in this study. Students' identities were masked before coding to ensure confidentiality.

The data were analyzed using a statistical software package, and means, standard deviations, and frequencies were calculated. The five exposure categories were recombined into three: never encountered, observed or assisted, and performed one or more times. This was done retrospectively so that the number of students in each category would be high enough for statistical comparison. Two one-way ANOVAs were run across these three categories of exposure, one to compare the students based on their confidence levels and the other to compare them based on their OSCE performances.

Results

The response rate for this study was 88% (148/172). Table 1 presents the results for level of exposure and confidence.

Nearly all students in both classes had taken a general adult history (99.3%) and a general pediatric history (97.3%). In fact, closer examination of the data showed that most students had performed these skills two or more times. The majority of the classes also had elicited, one or more times, a sexual history (96.7%), a history of drug or alcohol abuse (94.0%), and a history of sexual or physical abuse (59.6%). With respect to the higher-order communication skills, smaller proportions of the classes had performed these at all: breaking bad news to a patient or relative (50.7%),

101

Table 1. Levels of Exposure and Confidence in Communication Skills for the Dalhousie Medical School Graduating Classes, 1996 and 1997 (n = 148)

Communication Skill	Level of Exposure* Never Encountered (%)	Observed or Assisted (%)	Performed One or More Times (%)	Confidence† Mean (SD) (%)
Interviewing				
General adult history	.7	0	99.3	84.4 (12.2)
General pediatric history	1.3	1.3	97.3	76.8 (16.5)
Eliciting sexual history	.7	2.7	96.7	71.1 (19.5)
Eliciting history of drug or alcohol abuse	2.6	3.3	94.0	76.0 (17.4)
Eliciting history of sexual or physical abuse	21.2	19.2	59.6	53.4 (26.9)
Breaking bad news				
Breaking bad news (patient or relative)	6.0	43.3	50.7	51.7 (25.5)
Crisis management				
Managing a patient exhibiting drug-seeking behavior	16.2	45.9	37.8	50.7 (25.5)
Managing a violent or hostile patient	13.5	48.6	37.8	46.7 (25.8)
Counseling				
Providing counseling for drug or alcohol abuse	29.1	41.7	29.1	43.0 (26.8)
Providing counseling for victim of physical or sexual abuse	55.6	33.8	10.6	30.5 (24.5)

*Scale categories were collapsed to create this table as follows: "observed only" and "assisted senior staff member" were collapsed to "observed or assisted." "Performed once," and "performed two or more times" were collapsed to "performed one or more times."

†Distance marked along the 6-cm visual analog scale was converted to percentage of total length of scale.

managing a patient seeking drugs (37.8%), managing a violent or hostile patient (37.8%), counseling for drug or alcohol abuse (29.1%), and counseling for victims of physical or sexual abuse (10.6%).

The students in the graduating classes of 1996 and 1997 rated their confidence in interviewing relatively high for general adult history (84.4%), general pediatric history (76.8%), eliciting sexual history (71.1%), eliciting history of drug or alcohol abuse (76.0%), and eliciting history of sexual or physical abuse (53.4%). In the areas of breaking bad news and crisis management, ratings were around or below 50% (see Table 1). Lower confidence ratings were given to the counseling areas (i.e., drug and alcohol abuse (43.0%), physical or sexual abuse (30.5%).

Since the complexity of the higher-order skills may have contributed to lower confidence, seven individual skills were examined (see Table 2). The students in the graduating classes of 1996 and 1997 were more confident as their levels of exposure increased for each communication skill.

Confidence levels were higher for each of the seven skills examined for the group that had observed or assisted than they were for the group that had never encountered the skill. More dramatic differences were observed between the group that had performed the skill one or more times than for the group that had simply observed or assisted.

An ANOVA on the total score across the three OSCE communication stations (1997 class) was conducted for each of the

Table 2. Self-ratings of Confidence in Communication Skills by Levels by Exposure for the Dalhousie Medical School Graduating Classes, 1996 and 1997 (n = 148)*

Communication Skill	Level of Exposure Mean (SD) Never Encountered (%)	Observed or Assisted (%)	Performed One or More Times (%)	F-ratio†
Interviewing skills				
Eliciting history of drug or alcohol abuse	31.6 (11.3)	55.9 (14.1)	78.3 (15.2)	23.4
Eliciting history of sexual or physical abuse	28.4 (23.0)	38.5 (17.0)	67.3 (21.5)	48.7
Breaking bad news	23.4 (11.9)	36 9 (22.6)	67.9 (18.1)	50.3
Crisis management				
Managing a patient exhibiting drug-seeking behavior	14.6 (14.9)	46.5 (18.9)	71.5 (17.7)	89.6
Managing a violent or hostile patient	22.4 (22.2)	40.8 (20.8)	66.4 (19.0)	46.4
Counseling skills				
Providing counseing for drug and alcohol abuse	21.3 (19.8)	38.1 (18.2)	69.9 (20.8)	68.8
Providing counseling for victims of physical or sexual abuse	18.0 (17.1)	39.1 (20.8)	66.1 (22.4)	50.6

*Confidence ratings have been converted to percentage scores (0–100%). The first three interviewing skills are not included since nearly all students fell into the third group (performed one or more times).

†All F-ratios are statistically significant, $p < .001$.

following exposure groups: low exposure across all ten communication skills, medium exposure, and high exposure. The scores across the three stations were combined in order to achieve a more adequate representation of the students' performances. Since skills are context-dependent, this combined score yielded a more valid and reliable outcome measure. In order to provide a more defensible measure for the variable, exposure was defined as total exposure across all ten skills. We felt that a total exposure score would better represent students' actual medical school experiences in doctor–patient communication. The results showed that OSCE performances increased from the low-exposure group (n = 9; mean = 59.8) to the medium-exposure group (n = 80; mean = 64.3) to the high-exposure group (n = 58; mean = 66.3). These differences were statistically significant ($F = 3.1$; $p = .05$).

Discussion

In this study, graduating medical students had higher levels of exposure to standard communication skills than to higher-order communication skills, and their confidence levels were lower for the higher-order communication skills. One possible alternative explanation for the lower levels of confidence with respect to higher-order communication skills is that these skills are more demanding. Therefore, we decided to compare the confidence levels of students for each individual communication skill, as a function of type of exposure. The students who had performed each skill had much higher confidence levels than did those who had only observed or assisted. Also, the students who had observed or assisted with the skill had much higher confidence levels than did the students who had not encountered the skill at all. However, our findings suggests that observing or assisting is insufficient to develop confidence to an educationally significant degree; the more substantial gains were observed when students had performed the skill one or more times.

Although increased exposure increases confidence, a crucial question is whether increased exposure also leads to improved performance in applying these skills. The results of this study showed that this is indeed the case. The students who had had more overall exposure to the ten communication skills in this study performed at higher levels on the three OSCE stations emphasizing communication skills. Although not all ten communication skills were assessed in the three OSCE stations, these skills are composed of many common subskills, such as developing rapport, listening actively, explaining, and planning. Students with more exposure overall to the ten skills would have developed these subskills to a greater extent, and would most likely have better applied them in the OSCE.

It is important to note that students with less confidence in their abilities to exercise a skill may have avoided performing the skill in the clinical setting. Therefore, a causal relationship between exposure and confidence in this study should not be assumed. Although the results of the study confirmed our hypotheses, the exposure scale used did not measure actual level of exposure, i.e., *number of times* observed, assisted, or performed. Because the exposure scale simply measured students' recall of exposures, some bias may have been introduced. More important, the study surveyed only two medical school classes, and only one class's performance, so a broader survey is needed to confirm our findings.

The results of this study indicate that undergraduate students may not be getting sufficient opportunities to observe and practice complex communication skills in clinical or classroom settings, which results in low confidence levels. Factors affecting students' confidence do relate to clinical exposure, but they are also influenced by students' training in communication skills through structured programs that provide opportunities for learning and practice. The focus of this training appears to be on basic interviewing and interpersonal skills and not necessarily on higher-order skills. This was the case with the graduating classes in this study. All students had been given instruction in basic interviewing techniques. They had had opportunities to learn these techniques by observing videos and through role playing, by practicing their skills on each other and with simulated patients, and by receiving feedback on their skills from other students, course instructors, and simulated patients. For the higher-order skills, the students had been given exposure to breaking bad news through video programs and discussion as part of their training in palliative care; however, they had not had the opportunity to practice and receive feedback in these skills, as had been the case in their interviewing skills program. The students had been given no classroom training in crisis management or active counseling skills.

Both types of exposure may have to become more orchestrated for students during their undergraduate training. Providing effective training in higher-order communication skills as a core part of the undergraduate curriculum, where students have ongoing opportunities to observe, practice, and receive feedback in these skills, is a significant first step. This could occur in the clerkship years using the same techniques employed in learning basic interviewing skills. For this training, however, the use of role playing and standardized patients becomes particularly important. Once students have practiced these skills, they need to be provided with the opportunity to use them under supervision in a clinical setting. This practice will require some faculty development to ensure that physicians have the necessary skills to supervise effectively.

Correspondence: Dr. David M. Kaufman, Division of Medical Education, Clinical Research Centre, Room C-115, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H7.

References

1. Kauss DR, Robbins AS, Abrass I, Bakaitis RF, Anderson LA. The long term effectiveness of interpersonal training in medical school. J Med Educ. 1980;55:595–601.
2. Alroy G, Ber R, Kramer D. An evaluation of the short term effects of an interpersonal skills course. Med Educ. 1984;18:85–9.
3. Kendrick T, Freeling P. A communications skills course for preclinical students: evaluation of general practice based teaching using group methods. Med Educ. 1993;27:211–7.
4. Craig JL. Retention of interviewing skills learned by first-year medical students: a longitudinal study. Med Educ. 1992;26:276–81.
5. Engler CM, Saltzman GA, Walker ML, Wolf FM. Medical student acquisition and retention of communication and interviewing skills. Med Educ. 1981;56:572–9
6. Davis H, Nicholaou T. A comparison of the interviewing skills of first and final year medical students. Med Educ. 1992;26:441–7.
7. Evans BJ, Sweet B, Coman GJ. Behavioral assessment of the effectiveness of communication programs for medical students. Med Educ. 1993;27.344–50.
8. Dignan M, et al. Helping students respond to stressful interactions with cancer patients and their families: a pilot program. J Cancer Educ. 1989,4:179–83.
9. Bandura A. Self-efficacy mechanism in human agency. Am J Psychol. 1982;37:112–47.
10. Sanson-Fisher RW, Fairbairn S, Maguire P. Teaching skills in communication to medical students: a critical review of the methodology. Med Educ. 1981;15:33–7.

103

Assessment of Residents' Interpersonal Skills by Faculty Proctors and Standardized Patients: A Psychometric Analysis

MICHAEL B. DONNELLY, DAVID SLOAN, MARGARET PLYMALE, and RICHARD SCHWARTZ

The objective structured clinical examination (OSCE) has typically been found to be a valid and reliable method for assessing clinical knowledge and skills when evaluating performances of residents. For example, Sloan et al.[1] found a 19-problem, 38-station OSCE to be reliable (r_{xx} = .91) and valid in assessing the clinical skills of 56 surgical residents.

Often, OSCE performance is summarized in an overall score, which may represent a combination of history, physical examination, interpersonal and communication skills, technical skills, and organization. Interpersonal skills scores are sometimes reported separately because of their importance in overall performance. Warf et al.[2] found that when faculty judges evaluated general surgery residents' performances on a neurosurgical station there was no statistically significant difference between the junior and senior residents in performing the neurologic examination. Since general surgery residents do not receive training in neurosurgery beyond their intern year, it was not unexpected that there was no significant difference between levels of training. However, the senior residents were judged to be competent significantly more frequently than were the junior residents. It was also found that interpersonal skills correlated significantly with both competence and level of training. This study suggested that interpersonal skills are a very important facet of clinical competence that differentiates between residents at different skill levels.

Colliver et al.[3] also found statistically significant correlations (in the .30 to .50 range) between interpersonal skills and clinical competence. Similarly, Sloan et al.[4] found that global interpersonal skill judgments were moderately reliable and correlated highly with overall OSCE performance scores. Thus, it is clear that interpersonal skills are highly associated with the judged competency of medical students' or residents' performances.

Several studies have raised the question of who should evaluate interpersonal skills, a faculty proctor (FP) or the standardized patient (SP). Given the increasing clinical demands on faculty time, it is important to know whether SPs can assess interpersonal skills as validly and reliably as faculty members. Cooper and Mira[5] found that, on average, SPs gave more positive evaluations of communication skills of undergraduate medical students than did faculty members or other professional staff. They found that the communication scores derived from faculty's ratings did not correlate with the scores derived from the SPs' ratings.

Finlay et al.[6] assessed the communication skills of primary care physicians who had just received training in communication skills. Professional examiners and SPs evaluated the physicians' communication skills by means of a checklist. The two sets of scores correlated between .40 and .50 on the different OSCE problems, indicating that the SPs' evaluations cannot be used interchangeably with the faculty's evaluations.

In a test of the validity of eight faculty raters, Kalet et al.[7] videotaped the performances of 21 year-two medical students. Faculty evaluated the interviewing skills of those students on two different occasions using a checklist. The correlations of the communication scores among faculty members were low. Furthermore, the correlations between a faculty member's evaluations of the interviewing skills of the same students' performances on two occasions were also low.

A related question is whether checklist scores or global ratings provide more reliable and valid measures of performance. Regehr et al.[8] compared the psychometric properties of checklists with those of global rating scales on an eight-problem OSCE given to residents at all levels of training. They found better reliability and construct validity for global rating scales than for checklists. On the other hand, Hodges et al.[9] also evaluated the comparative reliability and validity of checklists and global ratings of communication skills. They found high correlations between global ratings and checklists.

Based on this review of the literature, we conclude that interpersonal skills are an important component of clinical competence. Global ratings are at least as valid and reliable as checklist scores. However, the levels of reliability and validity of interpersonal-skills ratings have not been clearly established. Also, it is not clear whether faculty or SPs, provide the more valid and reliable evaluations. The purpose of this study was to determine the psychometric characteristics of global interpersonal skills ratings of faculty proctors (FPs) and SPs.

Method

All 56 residents of a general surgery program participated in a 12-problem, 24-station surgery OSCE. Each OSCE problem was divided into two stations: Part A, in which a history and/or physical examination was performed or information was given to the SP, and Part B, in which the resident responded to several short-answer questions about the patient or SP seen in Part A. This study focused on the 12 Part A stations during which the FPs and SPs evaluated the residents' interpersonal skills.

Each of the 24 OSCE stations was five minutes in duration. At each station were either actual patients or SPs who had been trained to act in a consistent manner. As part of their training, they had been instructed in evaluating the residents' interpersonal skills. They had practiced making these evaluations during their training, formally evaluating the interpersonal skills of a resident, who was also evaluated by the trainer. Their evaluations were compared and the trainer and the SP discussed any differences in their evaluations. The typical training session lasted about one hour.

During each resident–patient encounter, an FP checked off indicated behaviors as they occurred. At the end of each of the Part A stations, the faculty member evaluated the resident's interpersonal skills (along with several other global performance dimensions). (Note that the trainer had reviewed the checklist and global ratings with each FP immediately before the OSCE.) Faculty rated their level of agreement with the statement "Interacted effectively with the patient" (0 = "not at all" to 4 = "very much"). The SPs independently evaluated the residents' interpersonal skills by telling the preceptors their ratings on the same five-point scale.

In order to determine the similarity of FPs' and SPs' ratings, the following analyses were done. First, the reliability of each of the 12 sets of paired FP and SP ratings was estimated by coefficient α. It was also estimated for the mean rating of the FP and the mean rating of the SP. Second, the reliability of the faculty's ratings of the residents' interpersonal skills across the 12 stations was estimated by means of coefficient α, and it was also calculated separately for the SPs' ratings. The Spearman–Brown formula was then used to estimate the expected reliabilities for two faculty raters and

TABLE 1. Psychometric Properties of Faculty Proctors' Ratings and Standardized Patients' Ratings of General Surgery Residents' Interpersonal Skills on a 12-Problem OSCE

Station	Reliability	Mean Difference	P-value Mean Difference	Faculty Construct Validity	Standardized Patient Construct Validity
Rating of faculty and patients (mean)	.92	−0.16	<.001	.68*	.73*
Neurosurgery	.94	−0.18	n.s.	.30*	.34*
Postoperative care	.85	−0.53	<.001	.35*	.32*
Plastics	.84	−0.14	n.s.	.55*	.59*
Breast options	.81	−0.42	<.001	.57*	.43*
Head and neck	.79	−0.16	n.s.	.41*	.56*
Breast examination	.72	0.00	n.s.	.39*	.19
Thyroid	.72	−0.30	<.003	.39*	.31*
Computed tomography	.70	0.05	n.s.	.10	.27*
Leg ulcer	.64	0.04	n.s.	.05	.31*
Abdomen history	.64	−0.04	n.s.	.49*	.36*
Biliary colic	.59	−0.52	<.001	.25	.44*
Hypercalcemia	.28	−0.06	n.s.	.31*	.21

*$p < .05$ (construct validity).

two SP raters. These Spearman–Brown estimates provide a standard against which to judge the magnitudes of paired FP and SP reliabilities.

It is possible to have relatively high reliabilities even though the FPs' and SPs' ratings may not be very closely calibrated. For example, an SP might be a more lenient evaluator than an FP. One indicator of similar calibration is that the mean rating of the FP is not significantly different from the corresponding mean rating of the SP. A two-way analysis of variance (faculty versus standardized patient, a "between-groups" factor; and comparing clinical problems, a "within-groups" factor) and analyses of simple effects were used to determine whether the FPs and the SPs evaluated the residents' interpersonal skills at approximately the same performance level.

If the paired FPs' and SPs' ratings are valid (convergent validity), they ought to correlate more highly with each other than with any other interpersonal skill rating. However, it is possible that this might not be the case. For example, faculty evaluations could correlate most highly with other faculty evaluations as SPs could with other SPs. To determine how the different ratings relate to one another, a hierarchical cluster analysis, using $1 - \text{Pearson } r$ as the similarity metric and the complete linkage amalgamation rule, was performed.[16] Clustering methods represent a variety of procedures that identify how variables group (cluster) together. Cluster analysis joins variables together based on the magnitude of the inter-correlations among the variables. A cluster is defined by two or more variables that correlate more highly with each other than they do with the other variables. We chose hierarchical cluster analysis over factor analysis because hierarchical cluster analysis better represents the relationships among variables when most of the variables intercorrelate substantially with each other. This analysis indicated whether the interpersonal-skills ratings clustered predominantly by (1) clinical problem (FP and SP couplets) or (2) FPs separately and SPs separately.

Finally, if the interpersonal-skills ratings are valid (construct validity), senior residents ought to perform better than junior residents and interns.[2] To this end, Pearson's correlations were calculated between interpersonal skills ratings and postgraduate year. These analyses were carried out for the 12 OSCE stations and the across—station averages for both FPs and SPs. Based on our experience with validity studies such as this, we expected the validity coefficients to be around .40 to .50. Fisher's z-test for differences between correlations was used to test whether the validity coefficients for the FPs and the SPs were significantly different from one another.

Results

The first data column of Table 1 presents the reliability coefficients for each of the paired (FP and SP) ratings for each of the 12 stations, and of the mean ratings of the FPs and SPs. The reliability of the mean FPs' and SPs' ratings is high, .92. The magnitudes of the reliabilities for the various stations vary: eight of these reliabilities were above .70, two were in the .60s, one was in the 50s, and one was .28.

The reliability of the faculty's interpersonal-skills ratings for the 12 OSCE stations was .77. The reliability was .74 for the SPs' ratings. The Spearman–Brown formula was used to estimate what these reliabilities would be if there were only two raters—to make them comparable to the paired (FP and SP) reliabilities. The estimated reliability of two faculty raters was .36, and it was .33 for two SPs.

If the ratings of FPs and SPs provide fairly equivalent information about the residents' interpersonal skills, there should not be significant differences in their mean ratings. The two-way analysis of variance comparing the equality of faculty's and SPs' means and the equality of means across problems indicated that (1) there was not a significant difference between the two groups ($p > .05$), (2) there were statistically significant differences among the means for the various OSCE problems ($p < .001$), and (3) there were significant interactions between groups and problems.

Since the significant interactions made the interpretation of the main effects equivocal, analyses of simple effects were performed to identify the exact pattern of differences. The second and third data columns of Table 1 summarize these analyses. As can be seen from this table, the differences in the mean FPs' ratings and the mean SPs' ratings (across the 12 stations) are statistically significant ($p < .001$). The mean difference is −0.16 on a five-point scale, indicating that the FPs tended to evaluate the residents' interpersonal skills at a slightly lower level than did the SPs. There were significant differences between paired ratings (FP and SP) for four of the OSCEs. In those four cases, the faculty evaluated the residents between three and five tenths of a scale point lower than did the SPs. In the other eight cases, the mean differences were small and not statistically significant.

If the interpersonal-skills ratings of a given FP and SP on a par-

105

ticular OSCE are valid, they should correlate highest with each other; however, they should also correlate significantly with the other measures if interpersonal skills is generally a valid construct. A hierarchical cluster analysis was performed to determine whether the FP and SP rating pairs for each station clustered closest. This dyadic clustering did take place for nine of the 12 possible OSCE stations. In these nine cases, the pairs correlated highest with each other. On two of the OSCE problems, the dyadic pairs did not correlate most highly with each other for unknown reasons, and on one OSCE problem, the pair did not cluster with each other because of the lack of variability in the SP's ratings.

To explore further the similarities in the ratings, the intercorrelations were calculated among the 24 different ratings of the residents' interpersonal skills. The median correlation among all 276 pairs of ratings was .20 (range −.24 to .89). The median correlation among the faculty's ratings was also .20 (range −.20 to .56), while the median was .17 (range −.16 to .50) for the SPs. In the case of the 12 paired correlations, the median correlation was .60 (range .20 to .89).

The construct validity of the FPs' and SPs' ratings was determined using the construct of experience; residents with greater experience should interact more effectively with patients than should junior residents. Pearson correlations were calculated between interpersonal skills ratings and level of experience. The fourth and fifth data columns of Table 1 present these correlations for the faculty and the SPs, respectively. None of the OSCE's paired correlations were significantly different from one another (Fisher's z-test for paired correlations). The average FP's rating (across the 12 stations) and the average SP's rating had higher construct validities than any of the individual interpersonal ratings. Construct validities of .68 and .73 are very high. In the experience of the authors, construct validities usually do not exceed .50. Nine of the 12 construct validities for the faculty ratings were statistically significant, while ten were significant for the SPs.

Discussion

In this study, faculty proctors and standardized patients were asked to evaluate residents' interpersonal skills at the end of each OSCE station. They made their judgments using a simple single-item scale. For the most part, the level of agreement (reliability) between the FPs and the SPs was adequate. On four of the stations, the reliabilities were sufficiently low to minimize their usefulness in making decisions about performance competency. On the other hand, the reliabilities of the average rating of the FPs and the average rating of the SPs were very satisfactory. Thus, it appears that these simple judgments are for the most part "reliable." On the other hand, the variability in the magnitudes of the reliability coefficients across stations suggests that one probably should not plan to make educational decisions about competency from performances at individual stations. Rather, it appears that one should use average performance measures.

An important consideration in estimating the reliability of ratings of interpersonal skills is whether to estimate reliability across problems or within problem pairs. The reliability of within-OSCE ratings is higher than that of between-OSCE ratings. It may be that interpersonal skills, like clinical reasoning skills, are affected by the context of the clinical case.

To explore this possibility, the 24 different ratings of the residents' interpersonal skills were intercorrelated. The median correlation among all possible combinations of raters, among the FPs and among the SPs was about .20. On the other hand, the median correlation among the paired ratings (FP and SP) was .60. Further, the hierarchical cluster analysis indicated that the interpersonal-skills ratings primarily clustered by OSCE station and not by rater type (FP or SP). This result has several implications. First, when the SPs and FPs are evaluating the same patient, their ratings tend to be more valid and more reliable than when the ratings are made on different OSCEs. The reliability appears to be more a function of the OSCE's case than the OSCE's evaluator. Standardized patients tend to give slightly higher evaluations than do faculty proctors. Our data do not indicate whether the FP or the SP is to be preferred.

In summary, global ratings of interpersonal skills are both reliable and valid. Faculty proctors and standardized patients appear to be interchangeable as evaluators of interpersonal skills. Case content is an important factor that influences residents' performances of interpersonal skills.

Correspondence: Michael B. Donnelly, PhD, Department of Surgery, C-243, University of Kentucky COM, 800 Rose Street, Lexington, KY 40536-0298.

References

1. Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The objective structured clinical examination: the new gold standard for evaluating postgraduate clinical performance. Ann Surg. 1995;222:735–42.
2. Warf BC, Donnelly MB, Schwartz RW, Sloan DA. The relative contributions of interpersonal and specific clinical skills to the perception of global clinical competence. J Surg Res. 1999;86:17–23.
3. Colliver JA, Swartz MH, Robbs RS, Cohen DS. Relationship between clinical competence and interpersonal and communication skills in standardized-patient assessment. Acad Med. 1999;74:271–4.
4. Sloan DA, Donnelly MB, Johnson SB, Schwartz RW, Strodel WE. Assessing surgical residents' and medical students' interpersonal skills. J Surg Res. 1994;57:613–8.
5. Cooper C, Mira M. Who should assess medical students' communication skills: their academic teachers or their patients? Med Educ. 1998;32:419–21.
6. Finlay IG, Stott NCH, Kinnersley P. The assessment of communication skills in palliative medicine: a comparison of the scores of examiners and simulated patients. Med Educ. 1995;29:424–9.
7. Kalet A, Earp JA, Kowlowitz V. How well do faculty evaluate the interviewing skills of medical students? J Gen Intern Med. 1992;7:499–505.
8. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med. 1998;73:993–7.
9. Hodges B, Turnbull J, Cohen R, Bienenstock A, Norman G. Evaluating communication skills in the objective structured clinical examination format: reliability and generalizability. Med Educ. 1996;30:38–43.
10. Dillon W, Goldstein M. Multivariate Analysis Methods and Applications. New York: John Wiley & Sons, 1984.

106

The Effects of Examiner Background, Station Organization, and Time of Exam on OSCE Scores Assessing Undergraduate Medical Students' Physical Examination Skills

CHRISTOPHER JAMES DOIG, PETER H. HARASYM, GORDON H. FICK, and JOHN S. BAUMBER

Since 1975, objective structured clinical examinations (OSCEs) have gained widespread acceptance as a method of making reliable assessments of clinical performance.[1] Standardized patients (SPs) function as patient, teacher, and evaluator by using their bodies as teaching and evaluation material. SPs can be asymptomatic, have stable findings, or be trained to simulate physical findings. SPs can be taught to portray a variety of standardized clinical presentations. Their participation in teaching and evaluating the complex clinical skills included in OSCEs has been well established.[2]

Research has demonstrated that multiple SP stations within the OSCE format may generate scores that vary greatly in reliability, from 0.20 to 0.95.[3,4] With large fluctuations in scores' reliabilities, research efforts have focused on the variables that can decrease or enhance the reliability of measurement. For example, inter-rater reliability was found not to be a deterrent to consistent measurement, and correlations generally varied from 0.80 to 0.90 between observers and raters when case-specific checklists were developed and if the items reflected observable behaviors. Due to the case-specificity phenomenon described by Elstein, many cases are generally needed to assess clinical competency within a defined problem (e.g., chest pain).[5] In other words, quality of performance on one case is a very poor predictor of performance on another.[6] However, if a single attribute is assessed, the number of cases required to attain reliable scores can be decreased (e.g., ten focused cases are required to assess the general skill of history taking, eight cases for physical examination, and 25 cases for differential diagnosis).[7]

Most OSCE stations employ a single case with a single SP and a single observer. However, because of the cost of OSCEs, efficiency would favor a station organized with two cases portrayed by a single SP. There are no research findings to indicate whether this organizational structure could adversely affect the reliability of measurement of an OSCE candidate's performance. Furthermore, OSCEs often use examiners from varied clinical backgrounds (e.g., residents, specialists, or family physicians). Given the importance of the OSCE's evaluation format and its predominant use for teaching and evaluating clinical skills, there is a need to determine whether the reliability of scores would be compromised by a rater's background, a station's organization, and the time of examination administration.

Method

Course Overview. The University of Calgary medical undergraduate program is three years in duration, with 11 instructional months per year. The first two years consist of "systems"-based courses using a problem-oriented curriculum that is taught in didactic lectures and small-group sessions. There is also a longitudinal medical skills course focusing on professional development and interdisciplinary skills, including a supervised setting for students to be instructed in physical examination. A "core document" given to each student provides detailed objectives for each physical examination maneuver. A standard physical examination textbook is recommended, and each student is provided with a six-hour video that shows local clinical experts demonstrating physical examination maneuvers. The instruction format is by small group. Preceptors are family physicians, specialists, or senior medicine residents. All small groups use SPs as instructional models. Further instruction in physical examination is carefully integrated into "clinical correlation" sessions within the systems courses. These sessions are organized so that the clinical correlation sessions build in an iterative fashion on skills learned in the sessions of the medical skills course. The instruction is also by small group. However, all preceptors are specialists within the area, and they provide patients as instructional models. These sessions expose students to clinical findings relative to each system and permit examination techniques to be observed and corrected by a clinical specialist within the area of study. At the end of the second year, the students take a certifying OSCE. the successful completion of which is a requirement for promotion into clinical clerkship (third year).

OSCE Station Development. The second-year OSCE consists of ten physical examination stations randomly selected from a bank of 44 stations. Each station tests one physical examination maneuver, and all were developed by one author (CJD) using the approach described. All maneuvers were selected from the core document's enabling objectives. Each maneuver was broken down into individual steps as outlined in the course's textbook. Each of these steps was identified as an item on a computerized examination score sheet. Criterion-based scoring was used, with each item scored as 0 (omitted or incorrect), 1 (partially correct), or 2 (correct).[8] Face and content validity of each checklist was established by review using a core group of physicians: five course preceptors, five medical educators with expertise in evaluation, five physicians with expertise in clinical teaching, and five specialists. The final content of each checklist and the minimum performance level (MPL) for each station were determined by consensus. It has previously been demonstrated that the validity of identifying the important items included in an OSCE station is superior when performed by a group of faculty compared with one individual.[9] Each station had been used in previous OSCEs, and the examination's properties established.

Examination Process. The medical skills examination included OSCE stations on history taking, physical examination, medical bioethics, and culture—health and illness. The examination totaled 3.5 hours, one hour of which was the physical examination section. The examination was conducted in one morning and one afternoon session. Each candidate completed ten physical examination maneuvers. At each station, there was one examiner and one standardized patient per pair of maneuvers. At each station, there was a short history to provide clinical context for each physical examination maneuver and students were given five minutes to demonstrate the first examination maneuver. The students then had one minute to review a short history for the second maneuver, and then five minutes to complete the second maneuver. At the end of 11 minutes, the students were given one minute to rotate to the next station (located in a separate examination room immediately adjacent to the preceding station) and to review the history for the first of the two maneuvers for the subsequent station. The physical layout of each station was standardized, with the patient dressed in appropriate examination apparel (but not draped or positioned for the examination), an examining tab. , necessary equipment on an adjacent table, and the examiner to one side.

Physical examination stations were grouped into two streams: Stream A paired maneuvers in one station that were from the same system or anatomic region, or that required a similar physical exam

107

TABLE 1. Summary of Individual Station and Overall Examination Results on a Ten-station Physical Examination Skills OSCE*

Examination Maneuver (No. Items per Checklist)	Student Examination Results			
	MPL†	Mean (SD)	Range	Proportion Successful
Ascites (10)	64.15	81.09 (16.59)	14.26–100	61/69
Cervical spine (13)	69.90	88.85 (11.10)	39.94–100	66/69
Jugular venous pulse (20)	67.68	74.64 (11.87)	35.45–100	55/69
Lung surface anatomy (14)	76.08	90.69 (11.87)	14.27–100	59/69
Median nerve (16)	53.78	61.08 (18.25)	13.84–100	50/69
Mini-mental status (18)	73.26	76.69 (9.07)	53.28–93.24	54/69
Peripheral arterial vasculature (17)	65.28	68.62 (16.96)	23.04–96.01	47/69
Shoulder (17)	61.47	72.60 (15.34)	38.42–100	54/69
Spleen (16)	57.06	75.53 (12.91)	23.78–95.10	63/69
Visual fields (14)	79.88	78.29 (16.22)	14.98–100	48/69
Overall (155)	66.72	75.93 (7.12)	56.95–91.37	65/69

*These are the ten physical examination maneuvers used during the examination. The maneuvers were paired into two streams of five stations. One stream paired maneuvers from similar body regions or physiologic systems (e.g., spleen and ascites), and one stream paired maneuvers from non-similar systems (e.g., shoulder and spleen).

†Minimum performance level or pass level.

skill; Stream B paired physical exam maneuvers that were not similar in region or skill examined. The pairings and sequence of examination maneuvers in Stream A were spleen and ascites, mini-mental status exam and median nerve, jugular venous pulse (JVP) and peripheral arterial system, shoulder and cervical spine, and visual fields and lung surface anatomy (the final pairing representing an understanding of clinical correlative anatomy). The pairings and sequence of examination maneuvers in Stream B were cervical spine and JVP, ascites and peripheral arterial system, lung surface anatomy and median nerve, mini-mental status and visual fields, and shoulder and spleen exams. Each stream ran in parallel during the morning and afternoon sessions. The pairings and examination maneuver sequences within the two streams remained constant between the morning and afternoon sessions.

Each examiner was a physical examination course preceptor. Two weeks prior to the exam, the examiners were sent the following station-specific information: a photocopy of the maneuver-specific objectives, a photocopy of the textbook describing the examination, and the station checklist. Each examiner was asked to review the appropriate section of the videotape (the videotape had been previously provided). An instructional session was held with all examiners to review the stations' expectations, checklists, and performance, and to discuss concerns. The examiners were not aware of the method of station validation, or the stations' minimum performance levels (MPLs). Six examiners were internal medicine residents, eight were family practitioners, and six were specialists.

An administrative assistant, unaware of the study's hypotheses, randomly allocated both the examiners and students to Streams 5 A and B, and times of examination (A.M. or P.M.).

Statistical Analysis. We hypothesized that the type of examiner, the stations' pairings of maneuvers that required similar content knowledge (extrapolated as being from the same examination systems), and times of examination would not contribute significant variance to the overall measure of examination reliability. For analysis, we used the general estimating equation (GEE) method, a modification of the generalized linear model (GLM).[10] GEE modeling is a robust and validated method of random-effects multivariate modeling that estimates general linear models but also permits a priori specification of a within-student correlation structure. In summary, the model provides an analysis of variance, but permits control of the potential effect of unequal distribution of data and the necessity to account for repeated measures. We used the exchangeable correlation structure within the GEE method to estimate the effects of the individual covariates (and any interactions) on the dependent variable of student performance.[10] As the sequence of examination maneuvers at each station was held constant within each stream, this was not included in the final analysis model, nor did we model within-examiner correlations. All analyses were performed with a statistical software package.

Results

Sixty-nine of 70 eligible students completed the examination: 35 were randomized to Stream A, and 34 to Stream B. The examination was structured, based on the availability of standardized patients, to have an unequal distribution between morning and afternoon sessions. Of the 69 students, 40 students were assigned to the morning examination, and 29 to the afternoon examination. Six examiners were residents, eight examiners were family physicians, and six examiners were specialists. The examiners were equally distributed between both streams and between morning and afternoon sessions.

The alpha coefficient for the examination was 0.84. The MPL for the examination was 66.85%, based on an equal weighting of the MPLs from the ten stations. Sixty-five of the 69 students were rated satisfactory on the overall physical skills examination. The mean performance was 76.81% ± 7.35 (mean ± SD). The range was from a low score of 56.51% to a high score of 92.28%. The performances at the individual stations are presented in Table 1. The overall mean score for candidates observed by senior internal medicine residents was 75.55%, that for candidates observed by family physicians was 79.22% (p = 0.07 compared with residents or specialists), and that for candidates observed by specialists was 75.28% (p = 0.38 compared with residents). No practical difference was observed in the candidates' performances by stream assignment: Stream A 77.00% and Stream B 76.61%. No practical difference was observed between the performances of candidates during the morning sessions (77.51%) and candidates during the afternoon sessions (76.00%). There was no within-stream between-examiner effect, and no within-time of examination between-examiner effect demonstrated. An unexplained difference was observed between the interaction of stream assignment and time of examination: morning session Stream A = 74.50%, Stream B = 80.52%, and afternoon session Stream A = 80.59% and Stream B = 71.40%. This observed interaction could not be explained by an effect of examiners. Given that the SPs and the pairings and sequences of examination maneuvers within the stations did not change, and in the absence of an alternate plausible explanation, the observed interaction was presumed to be due to a random effect of individual candidate performances.

101 108

Conclusions

Using a sound research design and robust analytic techniques, there was no evidence from this study that the variables—station organization, time of examination, and clinical background of examiner—contributed significant variance to the overall reliability of an OSCE assessing physical examination skills. With two parallel streams, and therefore two SPs' simulating the same physical examination maneuver, we assessed and found no difference in the between-SPs' mean value (form-within-case difference, as previously suggested by Battles[11]) for each physical examination maneuver (data not presented), which supports a conclusion that bias in our results was not introduced by the two SPs' demonstrating the same maneuver. Our assessment of only physical examination maneuvers is similar to the study of Kowlowitz and colleagues and that of Li and colleagues, and supports the reliability of our examination.[12,13] The difference in examiners' performances between family physicians and internal medicine residents or specialists did demonstrate a trend toward significance, and the absence of a statistically different result may have reflected a type II error. The effect of the examiner's background on rating students' performances requires further study.

Though OSCE examinations have gained widespread acceptance, major practical impediments remain in their cost and their labor-intensive organization. Reznick estimated the total costs for developing an OSCE and administering it to 120 students in a single medical school to be from a high of $104,400 to a low of $59,460, or $496 to $870 per student (Canadian denomination—CND).[14] For administering the exam only, costs ranged from $19,200 to $34,500 (CND) if examiners and SPs were paid, or from $16,500 to $19,200 (CND) if only SPs were paid (both estimates include catering costs for both examiners and SPs). In previous examinations using ten physical examination maneuvers, but without pairing of maneuvers within one station, we required 40 examiners and 40 standardized patients. The large numbers of examiners and SPs were a significant cost and administrative burden for our examinations, and they were important factors in our adopting the paired station strategy. In two previous examinations without paired stations, these examinations had an average alpha of .76. Our current study's findings support the premise that the pairing and sequencing of stations will not reduce the reliability of the assessment of a candidate's performance.

Reorganizing the assessment of physical examination skills within an OSCE by station by using maneuver pairing may contribute to improvement in overall efficiency and provide significant cost savings by reducing the numbers of SPs and examiners needed. Whether this can be applied in the assessment of other clinical skills in an OSCE requires further evaluation.

Correspondence: Dr. Christopher James Doig, Assistant Professor, Room EG23G, Foothills Medical Centre, 1403 29th Street NW, Calgary, AB, Canada T2N 2T9; e-mail (cdoig@ucalgary.ca).

References

1. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ. 1979;13:41–51.
2. Barrows H. Simulated Patients (Programmed Patients). Springfield, IL: Charles C Thomas, 1971.
3. Stillman P, Rutala P, Nicholson G, Sabers D, Stillman A. Measurement of clinical competence of residents using patient instructors. Proc Annu Conf Res Med Educ. 1982;21:111–6.
4. Harasym PH, Mohtadi NG, Henningsmoen H. The use of critical station to determine clinical competency in a "high stakes" OSCE. Scherpbier, van de Vleuten, Rethan (eds). In: Advances in Medical Education Dordrecht, The Netherlands: Kluwer Academic Publishers, 1997:661–4.
5. Elstein A, Shulman L, Sprafka S. Medical Problem Solving. Cambridge, MA: Harvard University Press, 1978.
6. Norman G, Tugwell P, Feightner J, Muzzin L, Jacoby L. Knowledge and clinical problem solving. Med Educ. 1985;19:344–56.
7. Stillman PL, Swanson DB, Smee S, Stillman A, Ebert TE, Emmel VS. Psychometric Characteristics of Standardized Patients for Assessment of Clinical Skills. Final Report support by American Board of Internal Medicine, January 1986.
8. Thompson WG, Lipkin M Jr, Gilbert DA, Guzzo RA, Robertson L. Assessment of the American Board of Internal Medicine resident evaluation form. J Gen Intern Med. 1990;5:214–7.
9. Valentino J, Donnelly MB, Sloan DA, Schwartz RW, Haydon RC. The reliability of six faculty members in identifying important OSCE items. Acad Med. 1998; 73:204–5.
10. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.
11. Battles JB, Carpenter JL, McIntire DD, Wagner JM. Analyzing and adjusting for variables in a large-scale standardized-patient examination. Acad Med. 1994;69: 370–6.
12. Kowlowitz V, Hoole AJ, Sloane PD. Implementing the objective structured clinical examination in a traditional medical school. Acad Med. 1991;66:345–57.
13. Li JTC. Assessment of basic physical examination skills of internal medicine residents. Acad Med. 1994;69:296–9.
14. Reznick RK, Smee S, Baumber JS, et al. Guidelines for estimating the real cost of an objective structured clinical examination. Acad Med. 1993;68:513–7.

109

Content, Culture, and Context: Determinants of Quality in Psychiatry Residency Programs

RACHEL YUDKOWSKY and ALAN SCHWARTZ

Residency training programs vary across characteristics such as their didactic and clinical experiences, attributes of the incoming residents, faculty characteristics, research conducted, community service performed by the program, and eventual practice choices of graduates. Which of these characteristics are most salient for evaluating the quality of a program?

Elliott[1] lists characteristics of graduates, cost–effectiveness, fair and ethical treatment of trainees, and meeting societal needs as important quality indicators. Iverson[2] takes a dimensional approach. His dimensions, with metrics, are: intake [U.S. Medical Licensing Examination (USMLE) scores of matched applicants]; customer satisfaction [percentage of available positions filled by match, and percentage filled by U.S. medical school graduates (USMGs)]; residency review committee (RRC) quintile scores; and outcome (specialty board pass rates). The Accreditation Council for Graduate Medical Education (ACGME) recently switched its focus from process variables to outcomes, and is encouraging RRCs to evaluate a program on how well it provides for six core competencies: patient care, clinical science; interpersonal skills and communication, professionalism, practice-based learning and improvement, and systems-based practice.[3]

In 1997, a task force of the American Association of Directors of Psychiatry Residency Training (AADPRT) developed a survey to define the variables important to determining a program's quality from the psychiatry resident's perspective. The 41-item questionnaire was based on feedback from focus groups of psychiatry residents and program directors and a review of the literature. A total of 180 psychiatry residents from 16 programs completed the survey. Quality of supervision and teaching conferences, respect of faculty for residents, responsiveness of the program to feedback from residents, and morale in the department were the items most important to residents' satisfaction. A detailed description of the construction of the survey and its results was published by Elliott.[4]

In 1998, the AADPRT's survey was repeated with psychiatry residency directors and heads of major rotations to see whether their values agreed with those of the residents.[5] This paper describes the use of multidimensional scaling (MDS) of the survey's results to establish whether there were distinct groupings of program directors with different opinions about the determinants of quality in psychiatry residency programs. These groupings might represent types of psychiatry programs (or market niches) as reflected in the values and priorities of their faculty and directors.

Method

Multidimensional scaling is an analytic technique frequently used in marketing research to identify the psychological dimensions underlying customers' preferences with respect to multiple variables or features of a product.[6] In MDS the difference between clusters or groups of variables is predicted by the distance between the variables in psychological "space," with the dimensionality of the space equal to the number of relevant dimensions underlying the data. These dimensions can be thought of as analogous to the latent constructs derived in factor analysis. The scaling algorithm derives the dimensions and plots the coordinates of the variables in the resulting multidimensional space. MDS is an inherently interpretive procedure—it locates variables on dimensions but requires the investigator to determine whether the dimensions can be intelligibly labeled.

Individual Differences Scaling (INDSCAL) is an MDS algorithm that models both the overall *dimensions* that underlie the perceptions of the group of respondents and individual *weights* on those dimensions for each respondent, allowing individuals to vary in the importance they attach to each of the dimensions. For example, for one individual, dimension A (the educational resources available, for example) may be highly salient, while dimension B (the administration of the program, for example) is relatively unimportant. For another individual, these priorities may be reversed. By examining the distribution of subject weights one can identify clusters of subjects who share similar values regarding the relative importances of the various dimensions.

The questionnaire was sent in late 1998 to all psychiatry residency directors listed in the American Medical Association's 1998–1999 Directory of Accredited Graduate Medical Education (GME) Programs. The faculty members who served as the heads of the inpatient and outpatient psychiatry rotations of each program were also surveyed. These are the two major rotations of psychiatry training programs, and the opinions of the heads of these rotations (henceforth referred to as service chiefs) would most likely represent the dominant values of the program.

The survey asked directors and service chiefs to rate how important the 41 items of the questionnaire were in determining the quality of a residency program. The anchors were 1 = least important, 4 = average importance, and 7 = most important.

Multidimensional scaling using INDSCAL was done on the survey responses. Solutions in two to six dimensions were generated.

Results

Of the 186 active programs listed in the GME directory, 117 programs (63%) responded to the survey. There were 234 individual responses from the 117 programs. Of these, 142 (61%) were from program directors and 92 (39%) were from service chiefs who were not identified as directors. For some programs the head of inpatient or outpatient services also served as an associate program director, confirming our supposition that these faculty members represent the administrative backbone of the program.

The Pearson correlation between the responses of the residency directors and those of the service chiefs was 0.98 ($p < 0.01$). We therefore pooled data from both chiefs' and residency directors' responses for the following analysis.

The two-dimensional INDSCAL solution was degenerate and was discarded. The solutions in three to six dimensions were inspected for goodness of fit and interpretability. The three-dimensional configuration provided the most interpretable dimensions, and accounted for 46.4% of the variance in the data; higher-dimensional solutions accounted for only slightly more variance.

Based on examination of the locations of the items, particularly those that had particularly high or low coordinates in each dimension, the dimensions seemed to correspond to three constructs: "curriculum," "quality of the institution," and "supportiveness of the administration of the program." The three dimensions, and the highest-loading items on each, are given in Table 1.

Subject weights measure the importance or salience of each dimension to each respondent; they range from 0 (completely ig-

Table 1. Three Dimensions of Quality of Psychiatry Residency Programs, Based on Multidimensional Scaling of Responses by Residency Directors and Service Chiefs to a 1998 Questionnaire*

Dimension	Questionnaire Items That Load Highly on the Dimension
Curriculum	Quality of supervision, training in biomedical and psychosocial psychiatry and the balance between them, diversity of patients and settings, opportunities for continuity of care, responsibility for patient care
Quality of the institution	Academic reputation of institution, clinical reputation of faculty, opportunities for research and teaching; board scores of graduates, job satisfaction of graduates
Supportiveness of the program administration	Fairness in evaluation of residents, respect of faculty towards residents, personal qualities and administrative abilities of the program director, responsiveness of the program to feedback from residents

*The questionnaire asked respondents to rate the importance of 41 items in determining the quality of psychiatry residency programs. A total of 234 program directors and service chiefs from 117 programs completed the questionnaire.

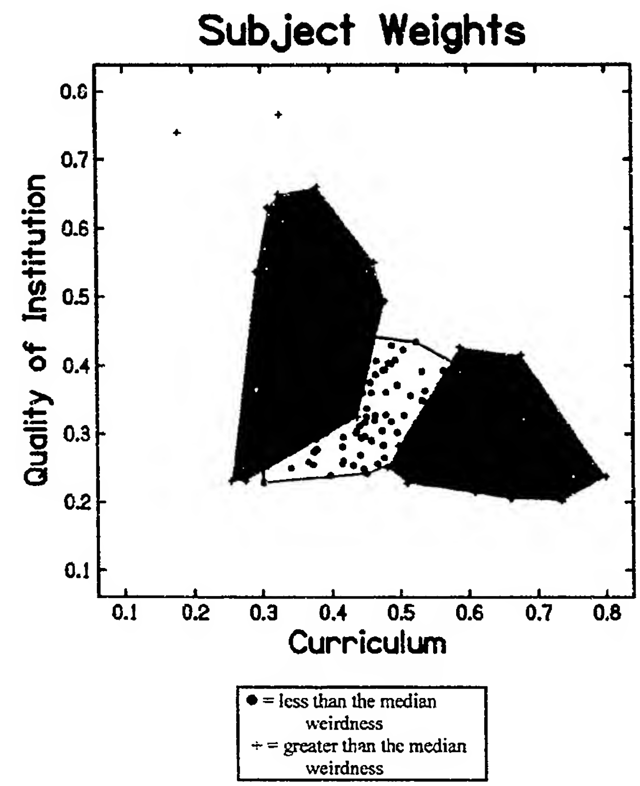

Figure 1. Subject weights for the "curriculum" and "quality of the institution" dimensions of the three-dimensional INDSCAL solution. The unshaded polygon encloses subjects with less than the median weirdness (based on weights in all three dimensions), represented as circles; the two shaded polygons identify two groups of subjects with greater than median weirdness, represented as crosses. Two subjects in the upper left had extreme (outlier) weirdness values. "Weirdness" measures how similar each respondent is to the typical respondent, based on the relative importance of each dimension and the goodness of fit for that respondent.

nored) to 1 (overwhelmingly important), and need not sum to 1. In addition, each respondent is assigned a "weirdness" value, which measures the similarity of his or her responses to those of the typical respondent, based on the relative importance of each dimension and the goodness of fit for that respondent.

There was a great deal of variation in individual preferences, but no distinct clusters were evident. Notably, although weights for the supportiveness of the administration of the program dimension fell between 0.25 and 0.40 for nearly all respondents, the importance attributed to the dimensions of curriculum and quality of the institution varied extensively across individuals. Figure 1 plots the weights of curriculum and quality of the institution against one another for each respondent. The unshaded polygon encloses data for more typical respondents with less than the median weirdness, represented as circles; the two shaded polygons identify two groups of less typical respondents with more than the median weirdness, represented as crosses. Two respondents in the upper left had extreme (outlier) weirdness values.

While the most typical respondents gave curriculum weights between 0.3 and 0.6, and quality of the institution weights between 0.2 and 0.4, two groups of respondents displayed different weight patterns. One group (lower right) gave curriculum substantially higher weights than typical (ranging from 0.5 to 0.75); the other group (upper left) gave quality of the institution substantially higher weights than typical (ranging from 0.25 to 0.65). On the average, most respondents' data displayed a continuum of weight patterns in which curriculum was considered to be more important than either quality of the institution or supportiveness of the administration of the program; the respondents with the lowest weirdness scores weighted these dimensions in the proportions 1.3:1:1, respectively.

Discussion

The three dimensions that emerged from the MDS are consistent with the many suggested quality indicators reviewed above.[1–3] How might we conceptualize this triad?

The dimensions of curriculum and supportiveness reflect two different aspects of the process of residency training. The curriculum dimension describes the *content* of the educational program; the supportiveness dimension reflects the *culture* or ambiance within which the training occurs. Residency directors and their faculty seem to differentiate between these two aspects of the program, and value both as indicators of the quality of the program. The dimension of institutional quality includes items reflecting the reputation and resources of the institution as well as items generally considered to be outcomes (i.e., board scores of graduates and graduates' job satisfaction). In this instance, these "outcome" items probably serve as proxies for (and a reflection of) the reputation of the institution and the quality of the residents it attracts, rather than as true outcome measures. This dimension could represent a general *context* factor, reflecting the quality of the facilities, the faculty, and of the residents themselves. This general factor could itself modify the effects of the process variables either up or down, thereby affecting the expected outcomes. Thus, this dimension may reflect the expectation of program directors and chiefs that equivalent processes (curriculum and supportiveness) could lead to better outcomes if they are provided in the context of a higher-quality institution.

Donabedian lists input, process, and product as the dimensions of quality in healthcare.[7] Interestingly, product (outcomes) did not emerge as a dimension of the quality of a program. Perhaps residency directors and service chiefs focus on process variables as indicators of quality since it is in the process of residency training that they deal. The neglect of outcome measures may also reflect a philosophy that, while the program is responsible for teaching, it is the residents' responsibility to learn. Outcome measures are highly confounded by the abilities and characteristics of the indi-

111

vidual residents and, thus, may not be considered an accurate or reliable measure of the quality of the program per se.

The ACGME and others who have begun the move towards outcome evaluation may wish to take this as a word of caution. Outcomes should not be considered in a vacuum. For at least some key stakeholders—the residents and faculty of the program—the context, content, and culture of the program are significant as well.

No truly distinct clusters or groups of respondents emerged from the multidimensional scaling of the data. While there may be programs with different missions—programs oriented towards research or community psychiatry, for example—these missions do not seem to result in drastically different definitions of quality. This suggests that there is a core concept of quality that holds across contexts and across missions—consistent with the RRC's model of minimum standards.

On the other hand, there seems to be a continuum of individual variation, rather than variation based on group membership. The individual respondents differed widely in the dimensions they considered most important. Since the individuals in this case are the faculty leaders and directors of the programs, these priorities are most likely reflected in the programs as a whole. The individual variation can be usefully segregated into three "market niches," corresponding to the three polygons in Figure 1. Thus we could describe three types of programs: (1) programs in which the quality of the sponsoring institution (context) is paramount, (2) programs in which the quality of the curriculum (content) is paramount, and (3) programs with a more typical weighting of the three dimensions.

While there was less variability in the importance attached to the supportiveness (culture) of the program, this should not be construed as lack of salience. Rather, *all* programs should be alert to the importance of this dimension.

Residents, too, vary in the levels of importance they attribute to the various features of a training program.[4] The context, content, and culture of a program may provide a good conceptual model of the dimensions along which the market varies. Programs may find it useful to identify and market themselves on the basis of these dimensions.

This study focused on only one of the many stakeholder groups of residency programs. The needs and expectations of other stakeholders, such as program funders and employers of a program's graduates, remain to be defined. This study also focused only on psychiatry programs, but the dimensions of context, content, and culture would seem to be potentially applicable to other specialties as well. Repeating the study with other specialties will tell us whether indeed this triad is relevant to the quality of programs across specialties.

The method of multidimensional scaling is a novel one for determining quality measures in graduate medical education. Repeated use of this technique, across stakeholders and across specialties, can help elucidate the factors most important to the evaluation of residency programs.

Correspondence: Rachel Yudkowsky, MD, Department of Medical Education, MC 591 UIC-COM, 808 South Wood Street, Chicago, IL 60612-7309; e-mail (rachely@uic.edu); Reprints are not available.

References

1. Elliott RL, Juthani NV, Rubin EH, Greenfield D, Skelton WD, Yudkowsky R. Quality in residency training: toward a broader, multidimensional definition. Acad Med. 1996;71:243–7.
2. Iverson DJ. Meritocracy in graduate medical education? Some suggestions for creating a report card. Acad Med. 1998;73:1223–5.
3. Accreditation Council for Graduate Medical Education. ACGME Outcomes Project: General Competencies. 9/28/1999 (http://www.acgme.org/outcomes/comptv13.htm). Accessed 5/25/00.
4. Elliott RL, Yudkowsky R, Vogel RL. Quality in psychiatric training. II: Development of a resident satisfaction questionnaire. Acad Psychiatry. 2000;24:41–6.
5. Yudkowsky R. What determines the quality of psychiatry residency programs? Opinions of program directors and faculty. Master's thesis, University of Illinois at Chicago, May 2000.
6. Carroll, JD, Chang, JJ: Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart–Young decomposition. Psychometrika. 1970;32:283–319.
7. Donabedian A. Evaluating the quality of medical care. Milbank Mem Fund Q. 1966;44(part 2):166–206.

112

Gauging the Outcomes of Change in a New Medical Curriculum: Students' Perceptions of Progress toward Educational Goals

GREGORY MAKOUL, RAYMOND H. CURRY, and JASON A. THOMPSON

After decades of concern about the lack of momentum in reforming medical curricula, a number of schools have introduced significant revisions and innovations in recent years. In most cases, the goals of these changes have followed the general principles promulgated by the Association of American Medical Colleges' (AAMC's) General Professional Education of the Physician (GPEP) and College Preparation for Medicine Report and other similar documents.[1,2] Objectives consistent with these goals have been codified and disseminated through the AAMC's Medical School Objectives Project (MSOP).[3] Several new educational strategies (e.g., problem-based learning) and course domains (e.g., courses in professional skills and perspectives) have become common elements of the resulting curricular initiatives at many medical schools.[4]

Given the need to track the effects and effectiveness of change in medical education programs,[5–8] Makoul developed the Student Perception Survey,[9] which focuses on how students view both the learning environment and their own learning experiences. It was first administered at Northwestern University Medical School in 1993, and has since been used by medical schools at the University of Chicago, Washington University, University of Utah, Medical University of South Carolina and, most recently, the University of Minnesota at Duluth. This study limits analysis to data collected at Northwestern between 1993 and 1999.

Context

In 1993, Northwestern University Medical School implemented a totally new first- and second-year (M1–M2) curriculum. Other, less sweeping, changes in the clinically oriented third- and fourth-year curriculum have been made more incrementally over the past decade, and are not a focus of this report. While some improvements have been made in our nearly seven years of experience with the M1–M2 curriculum, the basic concept and format are still firmly in place. The curriculum is composed of four courses, each presented in a series of discrete, topically focused units.[10] Each course and nearly every unit are interdisciplinary in nature and draw faculty from a number of departments; all are managed and funded centrally by the dean's administration.

Two areas of emphasis differentiate the current M1–M2 curriculum from its predecessor. The first is a change in the way we expect students to learn medicine. Our students are now explicitly regarded as adult learners, with a wide variety of backgrounds, aptitudes, and learning styles. Adult education models embrace this diversity and provide a framework for continuous self-directed education beyond the formal curriculum. Moreover, the very nature of the profession demands that students learn to "think on their feet," relating different areas of knowledge one to another and serving as critics of their own and others' reasoning processes. Accordingly, the curriculum provides a variety of learning formats, with an emphasis on interactive, discussion-based small-group activities. In addition, the clinical skills units include peer observation and feedback on a regular basis.[10,11]

The second emphasis is a dramatic increase in the attention paid to issues of professional perspectives and professional skills. As detailed by Curry and Makoul,[4] attention to students' interpersonal skills and attitudes and to the interface of the medical profession with society at large had grown steadily for some years. Not until the early 1990s, however, did schools begin to address these issues comprehensively. Since then, professionalism has become much more visible on the medical education agenda.[3] The conceptual framework of patient-centered medicine (also referred to as relationship-centered medicine), which highly values the physician's capacity for empathy, attentive listening, and concern for the patient's perspective,[12] has been instrumental in bringing about these changes.

The very breadth and comprehensiveness of significant educational reform make it difficult to reliably evaluate the specific impact of any component. Further, consistent with the focus on adult learning and professional development (i.e., we want our students to mature as self-aware professionals), we consider students' perceptions to be an important element of curriculum evaluation. We used the Student Perception Survey as our program evaluation tool because it offers a broad view of students' attitudes and experiences. For instance, we were interested in assessing, over a period of years, whether the new M1–M2 curriculum affected students' perceptions about the *importance* of key educational goals, and whether it had an effect on their perceived *progress* toward those goals.

Educational Goals: Importance. There is some concern that medical students become less idealistic and more cynical as they progress through the curriculum.[13,14] On the other hand, students are likely to place more emphasis on areas relevant to clinical practice as they approach the clinical clerkship phase of their education. To assess whether students place more or less value on key educational goals after their first two years of medical school, we can compare responses to the Student Perception Surveys administered to incoming students with those to surveys administered to the same students at the end of their second year (just before clinical clerkships begin). Since we expect that incoming students will highly value all of the goals, thus generating a ceiling effect, we do not expect the importance ratings to rise. Neither do we expect them to fall, since the new curriculum attempts to reinforce the value of these goals. Thus, our expectations regarding importance ratings are phrased as our first (null) hypothesis: There will be no statistically significant difference in the importance ratings when Student Perceptions Surveys administered to incoming students are compared with those administered at the end of the second year.

Educational Goals: Progress. Attending physicians' comments regarding the readiness and performances of students in their clerkships provide one good indication of whether a new M1–M2 curriculum is effective. However, it is difficult to systematically evaluate progress toward a variety of goals with such a method. Since we have a pass—fail grading system, the only grade-like metric available is the U.S. Medical Licensing Examination (USMLE) Step 1 score, also poorly suited to address a diverse set of goals. The Student Perception Survey allows us to assess students' views about the extent to which the curriculum has helped them progress toward each of the goals listed in Table 1. A brief "In Progress" article published in *Academic Medicine* reported immediate positive changes in ten of the 16 educational goals when data collected from the class of 1996, which progressed through the first two years before the curriculum was implemented, were compared with data from the classes of 1997 and 1998, the first cohorts to complete the new M1–M2 curriculum.[9] Since we expect the revised curriculum to prove effective in maintaining those changes, we offer the

113

TABLE 1. Responses to Importance of Educational Goals Section of the Student Perception Survey by Incoming and Experienced Students at Northwestern University Medical School, Classes of 1997–2001*

Educational Goal	Incoming Students Mean (SD)	Experienced Students Mean (SD)
Learn the language and information necessary for practicing medicine	3.90 (0.30)	3.85 (0.39)
Master skills for eliciting information from patients	3.84 (0.37)	3.84 (0.39)
Master physical examination skills	3.83 (0.40)	3.85 (0.39)
Become proficient in clinical decision making	3.86 (0.37)	3.76 (0.54)‡
Master skills for providing information to patients	3.73 (0.48)	3.71 (0.52)
Master skills for communication with colleagues	3.58 (0.55)	3.62 (0.58
Learn how to manage time more effectively§	3.30 (0.77)	3.34 (0.78)
Become more aware of ethical issues in medicine	3.38 (0.66)	3.25 (0.72)‡
Become more proficient at learning on your own	3.30 (0.79)	3.41 (0.74)†
Develop skills that will enhance lifelong learning	3.46 (0.69)	3.51 (0.67)
Develop skills for practicing health promotion and disease prevention	3.51 (0.65)	3.58 (0.61)
Understand how the stresses of life as a physician will affect your personal life§	3.18 (0.80)	3.24 (0.82)
Identify strengths and weaknesses in your academic and clinical abilities	3.51 (0.65)	3.52 (0.63)
Become more comfortable when being assessed by your peers§	3.09 (0.84)	3.05 (0.89)
Gain a full appreciation for political, economic, and social influences on health care§	3.18 (0.74)	3.20 (0.78)
Improve your problem-solving skills	3.49 (0.65)	3.63 (0.59)‡

*Medical students completed the Student Perception Survey at the start of their first year and again at the end of their second year. At each time point, the Educational Goals section of the survey asked them to rate the importance of the 16 goals reproduced in this table, using a five-point scale that runs from 0 ("not at all important") to 4 ("absolutely essential"), with the intervening points labeled as well. Paired *t*-tests were run to determine whether student perceptions of the goals changed significantly after two years in medical school (n = 511 pairs).

†$p < .01$, two-tailed.

‡$p < .001$, two-tailed.

§These goals were added by faculty who developed the new curriculum. The remainder are operationalizations of the original eight goals for medical education.[10]

second hypothesis: Students who have progressed through the new curriculum will report more progress toward educational goals than will students who completed the survey before the new curriculum was in place.

Method

Student Perception Survey. The survey gathers information about medical students' perceptions regarding faculty contact, educational goals, educational activities, and patient-centered tasks of care. It also gauges learning orientation, social orientation, career plan, conceptions of health, and demographic information. It is administered longitudinally via scan-form or computer: once at the beginning of medical school (i.e., during orientation week) and again at the end of the second year (i.e., just before clerkships). (We ran a study in 1997 to compare pencil-and-paper, scan-form, and computer versions of the survey; no difference in response patterns was detected.) This report includes data collected at both time points from students in the classes of 1996–2001. The survey is usually completed by all students in each cohort; it was distributed to fewer second-year students in 1995 and 1996, and fewer incoming students in 1998, due to administrative errors. Social security numbers serve as identification tags, allowing us to match surveys from incoming and experienced students without accessing their names or creating another set of identification numbers.

Educational Goals. In 1990, the dean, with the approval of all department chairs and senior deans, established eight goals for medical school education.[10] The 16 goals assessed in the Educational Goals section of the Student Perception Survey (see Table 1) were developed by explicating these original eight (e.g., operationalizing "communication") and then expanding the list to include four additional goals expressed by faculty who had developed the new curriculum for the first two years of medical school. Table 1 indicates which of the goals were added. Nunnally emphasized that the plan and procedure of an item's generation is a primary determinant of its content validity.[15] Drawing the items directly from goals outlined by the medical school certainly enhanced content validity. Further support comes from the observation that these goals are not unique to Northwestern; they are reflected in blueprints for medical education,[1-3] deemed relevant by the other schools using the Student Perception Survey, and in the expressed values of practicing physicians.[16] The items also have representational validity, as pilot tests conducted during the survey-development process indicated that medical students understood these items as intended.[17]

The Educational Goals section of the survey asks both incoming and experienced students to rate the importance of these 16 goals on a scale ranging from 0 = "not at all important" to 4 = "absolutely essential." The intervening scale points are labeled 1 = "slightly important," 2 = "moderately important," 3 = "very important." The survey administered at the end of the second year also asks students to indicate the extent to which their medical school experience has helped them progress toward each goal. The scale for measuring progress ranges from 0 = "not at all" to 4 = "completely."

- *Importance.* To test our first (null) hypothesis which posits little change in how students value the various educational goals, we performed paired *t*-tests on data from surveys administered to incoming and experienced students in the classes of 1997 through 2001, all of whom had been exposed to the new curriculum. Since we assert the null hypothesis, statistical power is an important consideration. Simply stated, the power of a test is the probability of rejecting the null hypothesis when it is indeed false. Given the large sample of matched pairs (n = 511), we chose a fairly conservative α level to avoid highlighting differences of trivial magnitude. At $\alpha = .01$ (two-tailed), we have statistical power greater than .98 for detecting small to medium effect sizes.[18]
- *Progress.* To test our second hypothesis, which states that the new curriculum should be associated with greater perceptions of progress toward the educational goals, we performed independent-sample *t*-tests on data from surveys administered to experienced students (those at the end of their second year). (One-way ANOVAs indicated that data from the classes of 1997 through 2001 could be combined because they were statistically similar. Thus, we ran *t*-tests to facilitate presentation and interpretation of results.) We compared the perceptions of students in the class of 1996 (n = 165), who had experienced the old curriculum, with those of students in the classes of 1997 through 2001 (n = 603). Again, the large sample size affords good statistical power. At $\alpha = .01$ (two-tailed), we have statistical power greater than .80 for detecting small to medium effect sizes via these independent-sample *t*-tests.[18]

Results

Importance. On average, the students rated all of the educational goals from "very important" to "absolutely essential" (see Table 1). When surveys administered at the two time points were matched

114

TABLE 2. **Experienced Students' Perceived Progress toward Educational Goals* While in the Old Curriculum (Class of 1996) Versus the New Curriculum (Classes of 1997–2001), Northwestern University Medical School**

Educational Goal	Old Curriculum (n = 165) Mean (SD)	New Curriculum (n = 603) Mean (SD)
Learn the language and information necessary for practicing medicine	2.67 (0.67)	2.86 (0.65)†
Master skills for eliciting information from patients	2.80 (0.65)	2.91 (0.69)
Master physical examination skills	2.58 (0.71)	2.48 (0.76)
Become proficient in clinical decision making	2.02 (0.83)	2.29 (0.81)†
Master skills for providing information to patients	1.81 (0.99)	2.31 (0.90)†
Master skills for communication with colleagues	2.25 (0.91)	2.52 (0.86)†
Learn how to manage time more effectively‡	2.19 (0.97)	2.26 (1.04)
Become more aware of ethical issues in medicine	2.62 (0.78)	2.88 (0.80)†
Become more proficient at learning on your own	2.60 (0.82)	2.93 (0.89)†
Develop skills that will enhance lifelong learning	2.47 (0.80)	2.75 (0.86)†
Develop skills for practicing health promotion and disease prevention	2.22 (0.78)	2.34 (0.90)
Understand how the stresses of life as a physician will affect your personal life‡	1.93 (1.03)	1.95 (1.07)
Identify strengths and weaknesses in your academic and clinical abilities	2.36 (0.84)	2.38 (0.88)
Become more comfortable when being assessed by your peers‡	1.92 (0.89)	2.27 (0.97)†
Gain a full appreciation for political, economic, and social influences on health care‡	1.82 (0.90)	2.24 (0.93)†
Improve your problem-solving skills	2.42 (0.77)	2.74 (0.78)†

*The Educational Goals section of the Student Perception Survey, distributed to students at the end of their second year, asked them to indicate the extent to which their medical school experience had helped them progress toward each of the 16 goals reproduced in this table, on a scale running from 0 ("not at all") to 4 ("completely"). Independent-sample t-tests were run to determine whether there were significant differences between the perceptions of students who had experienced the old curriculum (class of 1996) as compared with those who had experienced the new curriculum (classes of 1997–2001).

†$p < .001$, two-tailed.

‡These goals were added by faculty who developed the new curriculum. The remainder are operationalizations of the original eight goals for medical education.[10]

and importance ratings were compared via paired t-tests, we found statistically significant, though relatively small, differences (Δ) in how the students valued four educational goals. Importance ratings increased for "become more proficient at learning on your own" ($\Delta = .11$, $p < .01$) and "improve your problem solving skills" ($\Delta = .14$, $p < .001$); they decreased for "become proficient in clinical decision making" ($\Delta = -.10$, $p < .001$) and "become more aware of ethical issues in medicine" ($\Delta = -.13$, $p < .001$).

Progress. The students' mean ratings of the extent to which their experiences had helped them accomplish each goal were closer to the scale's mid-point that were the importance ratings (see Table 2). Students completing the new M1–M2 curriculum reported significantly more progress toward ten of the educational goals than did the cohort that progressed through the first two years before the new curriculum was implemented. The biggest changes were associated with "master skills for providing information to patients" ($\Delta = .50$, $p < .001$), "gain a full appreciation for political, economic, and social influences on health care" ($\Delta = .42$, $p < .001$), "become more comfortable when being assessed by your peers" ($\Delta = .35$, $p < .001$), "become more proficient at learning on your own" ($\Delta = .33$, $p < .001$), and "improve your problem solving skills" ($\Delta = .32$, $p < .001$). The only decrease was associated with "master physical examination skills ($\Delta = -.10$, ns).

Since distributions for some of the importance and progress items were not normal, we also ran nonparametric tests (Wilcoxon signed-ranks test for importance, Wilcoxon–Mann–Whitney test for progress). The power efficiencies of these tests are about 95% when compared with their parametric counterparts.[19] We obtained exactly the same patterns of statistical significance, reinforcing the notion that parametric tests are robust when it comes to the assumptions of normality.[20]

Discussion

A number of measures and methods (e.g., written tests, clinical skills exams, faculty reports) can provide data for assessment of students and curriculum evaluation. However, such data are relatively particular in nature. Just as clinical outcomes researchers obtain patients' perceptions to complement more objective measures of health,[21] medical educators interested in the outcomes of curricular reform have gained important information by measuring students' perceptions in the areas of well-being,[22] learning activity,[23] learning environment,[24] and long-term effects.[25] This study's findings indicate the value of gauging students' perceprions regarding a variety of education goals as well.

While there were statistically significant differences in importance ratings for 25% of the educational goals, there was no trend in terms of directionality. Thus, our first (null) hypothesis received general support; the value students placed on the educational goals remained relatively stable between orientarion week and the end of the second year of the curriculum. As shown in Table 2, our second hypothesis, which focused on progress estimates, received general support as well. Students who had progressed through the new curriculum reported more progress toward ten of the educational goals than did students who completed the survey before the new curriculum was in place. All of the statistically significant differences in progress estimates were larger than any of the differences in importance ratings. This pattern of results was immediate[9] and has been sustained over the years. It appears that the Patient, Physician & Society (PPS) course, which extends throughout the first two years,[10] contributes to increases in the students' perceived progress toward their educational goals. More specifically, the PPS course emphasizes providing information to patients, incorporates peer assessment and feedback, and explores the political, economic, and social influences on health care.

We were pleased to find that, when compared with the students in the old curriculum, the students who had experienced the current M1–M2 curriculum reported more perceived progress toward the goals of becoming more proficient at learning on their own and developing skills to enhance lifelong learning. We attribute this change to the adult-learner and active-learning approach taken by all four of the M1–M2 courses. However, we did not see a similar gain in the area of identifying strengths and weaknesses in academic and clinical abilities, an important component of lifelong learning and mindful practice.[26] The results suggest that we also need to focus our attention on helping students learn to manage time more effectively and understand how the stresses of life as a physician will affect their personal lives, two goals voiced by faculty who developed the new M1–M2 curriculum.

Regarding the goal of developing skills for practicing health promotion and disease prevention, we are planning to move to a more clinically oriented PPS unit on health risks, in part because the students reported little increased progress in this area at the end of their second year. Finally, despite a well-received first-year unit on physical examination skills in PPS, we observed a decrease in per-

115

ceived progress toward this skill set, a consistent and rather troubling finding over the years. We will continue to work toward improving students' confidence and competence in physical exam skills within the PPS course, as the first and second years of medical school offer an opportunity to ensure a consistent approach to teaching and learning basic skills. Our aim is to provide a solid foundation that can be built upon during the clerkships.

While it would have been preferable to collect the Student Perception Survey's data for more than one cohort in the old curriculum, the survey could not be implemented until it was designed and tested. Still, the pattern of results is clear and consistent, and changes in progress estimates can be logically linked to changes in the curriculum. Further, results from other schools using the Student Perception Survey reinforce the findings regarding progress. For instance, progress estimates also increased at the University of Utah after a curricular revision. Interestingly, significant progress toward a similar number of goals was evident at both Northwestern and Utah, but the pattern of results (i.e., mix and magnitude of changes) differed. (We will be working with Dr. Neal Whitman and colleagues at Utah to determine the extent to which observed changes reflect the emphases of M1–M2 curricular reform at that institution.) Students' perceived progress toward their educational goals did not increase at schools that did not make substantial changes in their M1–M2 curricula during the period they have used the Student Perception Survey.

Taken together, these observations highlight the generalizability and sensitivity of this approach to curriculum evaluation. The Student Perception Survey has proved a very useful tool for gauging the effects of curricular reform and identifying areas in need of more attention. We consider students' perceptions one important component of curriculum evaluation,[27] and we will continue to monitor them carefully. At present, we are working to develop a questionnaire for residency program directors and another one for medical school alumni, each of which will draw on aspects of the Student Perception Survey. As noted by Gerrity and Mahaffy,[5] this type of outcome data serves the important function of documenting where we have been and helping us better understand where we are going.

The authors thank Heather Sherman for her assistance in reviewing the literature related to this study and the RIME committee for helpful comments regarding the manuscript.

Correspondence: Gregory Makoul, PhD, Associate Professor and Director, Program in Communication & Medicine, Northwestern University Medical School, 750 North Lake Shore Drive (ABA 625), Chicago, IL 60611; e-mail (makoul@northwestern.edu).

References

1. Muller S (chair). Physicians for the Twenty-First Century: Report of the Project Panel on the General Professional Education of the Physician and College Preparation for Medicine. J Med Educ. 1984;59(11, part 2):1–208.
2. Education Committee, General Medical Council. Tomorrow's Doctors: Recommendations on Undergraduate Medical Education. London, U.K.: General Medical Council, 1993.
3. Association of American Medical Colleges. Learning Objectives for Medical Student Education: Guidelines for Medical Schools (Medical School Objectives Project Report I). Washington, DC: Association of American Medical Colleges, 1998.
4. Curry RH, Makoul G. The evolution of courses in professional skills and perspectives for medical students. Acad Med. 1998;73:10–3.
5. Gerrity M, Mahaffy J. Evaluating change in medical school curricula: how did we know where we were going. Acad Med. 1998;73:S55–S59.
6. Stone S, Qualters D. Course-based assessment: Implementing outcome assessment in medical education. Acad Med. 1998;73:397–401.
7. Friedman C, de Bliek R, Greer D, et al. Charting the winds of change: evaluating innovative medical curricula. Acad Med. 1990;65:8–14.
8. Rollins L, Lynch D, Owen J, Shipengrover J, Peel M, Chakravarthi S. Moving from policy to practice in curriculum change at the University of Virginia School of Medicine, East Carolina University School of Medicine, and SUNY–Buffalo School of Medicine. Acad Med. 1999;74(1 suppl)S104–S111.
9. Makoul G, Winter RJ. The student perception survey: a tool for assessing medical school curricula. Acad Med. 1997;72:410–1.
10. Makoul G, Curry RH. Patient, physician & society: Northwestern University Medical School. Acad Med. 1998;73:14–24.
11. Curry RH, Makoul G. An active learning approach to basic clinical skills. Acad Med. 1996;71:41–4.
12. Stewart M, Belle Brown J, Weston WW, McWhinney IR, McWilliam CL, Freeman TR. Patient-Centered Medicine: Transforming the Clinical Method. Thousand Oaks, CA: Sage Publications, 1995.
13. Becker HS, Geer B, Hughes EC, Strauss AL. Boys in White: Student Culture in Medical School. Chicago, IL: University of Chicago Press, 1961.
14. Sinclair S. Making Doctors: An Institutional Apprenticeship. New York: Berg, 1997.
15. Nunnally J. Psychometric Theory. New York: McGraw–Hill, 1978.
16. Shugars DA, O'Neil EH, Bader JD. Survey of Practitioners' Perceptions of Their Education. Durham, NC: The Pew Health Professions Commission, 1991.
17. Folger JP, Hewes DE, Poole MS. Coding social interaction. In: Dervin B, Voight MJ (eds). Progress in Communication Sciences. Vol IV. Norwood, NJ: Ablex, 1984:115–61.
18. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
19. Seigel S, Castellan NJ. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw–Hill, 1988.
20. Hanushek EA, Jackson JE. Statistical Methods for Social Scientists. New York: Academic Press, 1977.
21. Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. JAMA. 1999;282:1157–62.
22. Strayhorn G. Effect of a major curriculum revision on students' perceptions of well-being. Acad Med. 1989;64:25–9.
23. Blumberg P, Daugherty SR. A comparison of the perceived effectiveness of two educational methods in achieving school-related and practice-related goals. Teach Learn Med. 1994;6:86–90.
24. Robins LS, Alexander GL, Oh MS, Davis WK, Fantone JC. Effect of curricular change on student perceptions of the learning environment. Teach Learn Med. 1996;8:217–22.
25. Peters AS, Greenberger-Rosovsky R, Crowder C, Block SD, Moore GT. Long-term outcomes of the New Pathway Program at Harvard Medical School: a randomized controlled trial. Acad Med. 2000;75:470–9.
26. Epstein R. Mindful practice. JAMA. 1999;282:833–9.
27. Kern DE, Thomas PA, Howard DM, Bass EB. Curriculum Development for Medical Education: A Six-Step Approach. Baltimore, MD: The Johns Hopkins University Press, 1998.

An Index of Students' Satisfaction with Instruction

JAY H. SHORES, MICHAEL CLEARFIELD, and JERRY ALEXANDER

The purpose of this study was to determine whether a students' satisfaction index (SSI), derived from responses to a single rating of a faculty member's overall instructional ability, is a reliably valid tool for identifying those medical school faculty members whose instruction is in need of improvement.

Background

Debates have been held since the 1950s on the validity of students' evaluations of faculty (SEF).[1–6] While opinion is split over the application of SEF to the management of the professoriate,[7–9] the majority of the researchers addressing this issue support the use of SEF in the areas of faculty development and instructional improvement.[10–12]

The primary premises underlying use of students' ratings of the instructional abilities of faculty has been repeatedly addressed by educational researchers. The reliability and validity of such measures have been the subject of numerous studies.[1,3,14,15] While their results have not been consistent, they have predominantly supported the stability and validity of students' evaluations of faculty. The use of a single global measure to assess instructional ability has also received the attention of the researchers, and it has been established that such a measure can be valid.[18–20]

Students evaluate every instructor at the Texas College of Osteopathic Medicine. The items on the required evaluation form can be changed to correspond to the needs of academic departments and the unique characters of instructional programs. However, the final item on each and every evaluation form is constant across all courses. That item states "Overall, this is an effective instructor." Responses to this question are made on a five-point scale and reported on a 80-point scale (strongly agree equates to an SSI of 100, agree = 80, neutral = 60, disagree = 40, and strongly disagree = 20).

In practice, the SSI works like grades assigned to students in graduate and professional training programs. Medical students seldom use the lower half of the rating scale. Their responses result in SSI scores from the 60s through the 90s, with a mean score of 77. The SSI score is transmitted, along with responses to the other concepts assessed, to the faculty member and the department chair. Predictably, when an SSI drops below 70 the department's chair begins to comb through other data on teaching performance to find out why the faculty member is not doing well.

During the past 15 years faculty members have expressed a variety of concerns about the SSI. Faculty members think the SSI reflects the students' moods at the moment and, thus provides unreliable results. They feel that students are not capable of judging their instruction and that their peers would provide different (presumably more favorable) results. They suggest that the students' ratings correlate highly with the grades the students receive. Many feel that an assessment made at the end of the course disadvantages those who teach early. Each of these beliefs casts doubt on the reliability and validity of the SSI. These are the questions that are addressed in this study: (1) Is the SSI reliable? (2) Is the SSI valid? (3) Is the SSI biased by grades? (4) Is the SSI biased by the time lag between performance and measurement?

Method

Eighty-five second-year medical students in a five-month course in internal medicine agreed to evaluate quality of instruction for each of the 124 lectures given during the course. The 24 faculty teaching the course agreed to end their classes ten minutes early each day to allow the students time for evaluation. The departments of Medicine and Medical Education agreed to have faculty members present to evaluate each lecture. As an inducement for participation, the Department of Medical Education generated individual formative assessments and suggestions to improve the quality of instruction for each faculty member who taught for three or more hours in the course. These reports were delivered after all the data in the study had been collected.

Data were collected following each of 124 lectures from the entire portion of the 85 students who attended the lecture. Faculty also evaluated each of the lectures in the course. Complete data were collected for 24 instructors. One instructor was removed from the study due to the onset of acute illness during instruction. The 23 remaining instructors were each assessed by ten to 310 students, and by two to nine faculty members. One of the faculty evaluators in each session was from the Department of Medical Education; the rest were from teaching faculty in the Department of Medicine. At the end of the course the normal instructor-evaluation process was conducted and a post-course SSI was derived at that time.

Results

Is the SSI Reliable? To answer this question a test–retest was conducted. Students' evaluations at the end of the course were compared with those obtained from the same students at individual lectures. In Figure 1, faculty have been ranked by their end-of-course evaluations. Figure 1 presents a mean end-of-lecture SSI and a mean end-of-course SSI for each of the 23 faculty members. The end-of-lecture SSI is the average response of the students who attended the lectures presented by the instructor. The end-of-course SSI is the average response of the students who evaluated that instructor at the end of the course. The correlation between these two sets of satisfaction indexes is $r = .847$.

There is some difference between the means (end-of-lecture mean = 83.21, end-of-course mean = 82.72). A paired t-test of the difference between the means of the measures was not significant ($p = .647$). The practical significance of observed differences depends on their interpretation. One pair of observed differences could cause an instructor to be viewed as a member of a different group. Students' satisfaction with Instructor 2 shifted from moderate agreement to neutral. In the other 22 cases the instructors' ratings remained in the same relative positions on the criterion scale.

Is the SSI Valid? A concurrent validation was performed. Students' SSI scores were compared with those derived from the observations of the faculty members. Figure 2 presents end-of-lecture ratings from students and faculty for each of the 23 instructors. The students' end-of-lecture satisfaction index is the mean response of students who attended the lectures presented by the instructor. The faculty's end-of-lecture satisfaction index is the mean response from faculty members who attended the same lectures. The correlation between these two sets of satisfaction indexes is $r = .846$.

The measures had different mean values (students = 83.21, faculty = 79.23). A paired t-test of the difference between the means of the measures was statistically significant ($p = .012$). These instructors were consistently rated almost four points lower by faculty

117

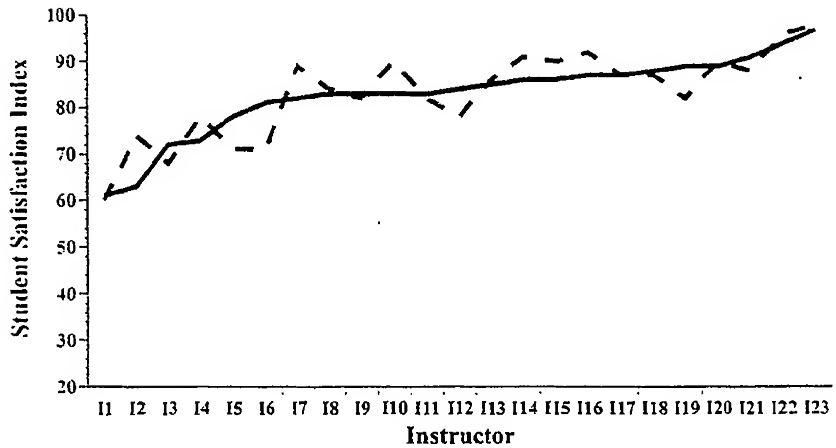

Figure 1. Students' ratings of 23 faculty immediately following each faculty member's lecture (broken line) and again at the end of the course (solid line). The faculty all delivered lectures at different points during the five-month-long internal medicine course.

than they were rated by students. On a practical basis, several pairs of these observed differences could cause an instructor to be viewed as a member of a different group. Instructors 3, 6, 10, 15, and 19 could be viewed as less competent on the basis of their peers' ratings.

Is the SSI Biased by Grades? Four non-cumulative examinations were given in the course. The grades received on the four exams were correlated with the average satisfaction index given by each student during the time block covered by each exam (n = 84). The resultant correlations were: $r = -.01$, .025, .089, and .090. There was no systematic relationship between grades and SSI scores.

Is the SSI Biased by Lag Time to Evaluation? End-of-course evaluations of the 23 instructors in this study followed their lectures by as much as five months. To assess the effect of lag time on the evaluations, each instructor was assigned to one of four groups based on the number of months that passed between his or her last lecture and the end-of-course evaluation. Table 1 presents descriptive data for the resultant groups. By subtracting the end-of-course satisfaction index from that obtained following their lectures, difference scores were generated. A one-way fixed-effects analysis of variance of the difference scores compared across groups was not significant (p = .821).

Incidental Findings. There is a ceiling effect in medical students' responses to the question that forms the basis for the satisfaction index. As a result, for both the end-of-lecture SSI and the end-of-course SSI there was less than one standard deviation available above the mean response in 52% of the cases. This reflects the fact that distributions of SSI responses were negatively skewed (−1.289).

The behaviors of the medical students changed during the study. They came to the lectures in far greater numbers than they had before. In fact, a secondary issue the authors wanted to investigate regarding the effect of attendance on performance could not be addressed. There were too few students in the "non-attending" pool to do an analysis.

Conclusions

The SSI demonstrated sufficient reliability and validity in this population of students and faculty to support its use by our institution as a tool to identify faculty members whose instruction is questionable. The measure did not appear to be biased by either earned course grades or lag time to evaluation. The marked negative skew and ceiling effect observed in the data limit the application of the SSI to its intended purpose, flagging poorly-performing instructors. Differentiation among the instructors at the upper end of the SSI is practically impossible.

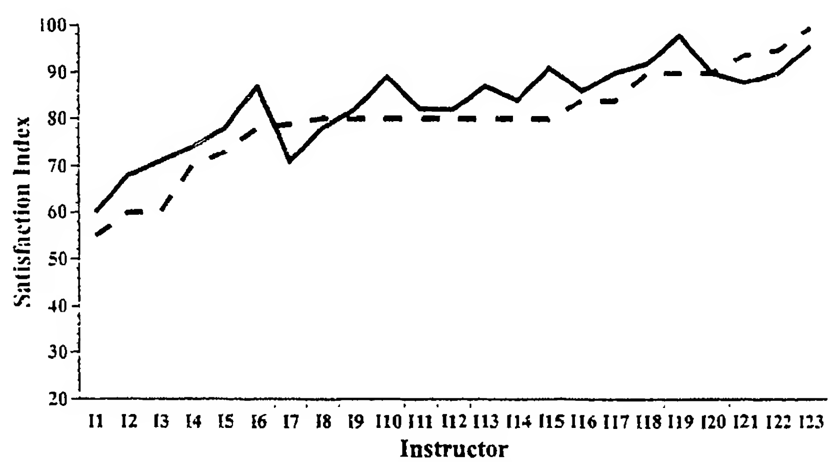

Figure 2. Students' (solid line) and faculty's (broken line) ratings of 23 instructors immediately following the lectures they delivered during a five-month-long internal medicine course.

118

TABLE 1. Students' End-of-course Satisfaction Indexes by Months of Lag from Lecture to Evaluation*

Lag Time from Lecture to End of Course	Satisfaction Index		
	Mean	SD	No.
≤1 month	82.33	9.64	9
2 months	83.10	15.40	4
3 months	84.35	3.33	4
≥4 months	81.97	5.57	6
TOTAL	82.72	8.68	23

*A total of 23 instructors gave 124 lectures during the five-month internal medicine course and all were evaluated by students in an end-of-course evaluation.

Discussion

The findings of this study support the assumption that a single item can be used to assess the global effectiveness of a faculty member's instruction.[17] However, caution should be used in generalizing the findings. The strength that the student satisfaction index has demonstrated may have been due, in part, to its use with second-year medical students. Medical students are highly focused intelligent respondents whose backgrounds are academically homogeneous. The strength of the index is also partially attributable to the fact that the respondents commonly used only half of the scale's values. This keeps the satisfaction index values for most instructors in a relatively narrow band and gives rise to a pronounced negative skew in their distribution.

The concurrent validity of the SSI was assessed by comparing students' responses with the responses of teaching faculty from the Department of Medicine and PhD-level medical educators. The study was conducted in a medical school that routinely evaluates every faculty member in every course. Data were collected and analyzed by a department that the students had come to trust for its grading of their examinations and forwarding of their assessments of faculty and courses to the administration. Both the students and the faculty knew they were engaged in a study of the SSI. The extent to which these environmental and research variables affected the results of this study will be known only when researchers subject the SSI to further analysis.

Finally, an ethical issue may be raised by the SSI. The SSI was developed as a flag to inform the faculty member and department chair that there could be a problem in the area of instruction. As such, it does not tell the user why the respondents feel that the instruction they have received is suspect; neither does it give guidance to assist in remedying the problem. It is incumbent upon an institution to develop the means for both accurately identifying the nature of the problem[18,19] and also addressing it productively[20-24] before it elects to discover which of its faculty has the problem.

Correspondence: Jay H. Shores, PhD, Department of Medical Education, University of North Texas HSC, 3500 Camp Bowie Boulevard, Forth Worth, TX 76107.

References

1. Cashin WE. Student ratings of teaching: The Research Revisited. (IDEA Paper No. 32). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development, 1995.
2. Cashin WE. Student Ratings of Teaching: A Summary of Research. (IDEA paper No. 20). Manhattan, KS: Kansas State University, Division of Continuing Education, 1988.
3. Marsh HW. Student evaluations of university teaching: research findings, methodological issues and directions for future research. Int J Educ Res. 1987;11:252–388.
4. Marsh HW. Multidimensional students' evaluations of teaching effectiveness; a test of alternative higher order structures. J Educ Psychol. 1991;83:285–96.
5. Marsh HW. A multidimensional perspective on students' evaluations of teaching effectiveness: reply to Abrami and d'Apollonia. J Educ Psychol. 1991;83:642.
6. Frey P. Validity of student instructional ratings: does timing matter? J Higher Educ. 1976;47:323–37.
7. Haskell RE. Academic freedom, tenure, and student evaluation of faculty. Educ Policy Analysis Arch. 1997;5(6): (http://olam.ed.asu.edu/epaa.html).
8. Cohen PA. Comment on a selective review of the validity of student ratings of teaching. J Higher Educ. 1983;54:448–58.
9. Feldman KA. Grades and college students' evaluations of their courses and teachers. Res Higher Educ. 1976;4:69–111.
10. Theall M, Franklin JL. Student Ratings of Instruction: Issues for Improving Practice. New Directions for Teaching and Learning, San Francisco, CA: Jossey-Bass, 1990.
11. Scriven, M. A unified theory approach to teacher evaluation. Stud Educ Eval. 1993;21:111–29.
12. Cashin WE, Downey RG. Using global student rating items for summative evaluation. J Educ Psychol. 1992;84:563–72.
13. Frey P. Validity of student instructional ratings, does timing matter? J Higher Educ. 1976;47:323–37.
14. Cohen PA. Comment on a selective review of the validity of student ratings of teaching. J Higher Educ. 1983;54:448–58.
15. Seldin P. Changing Practices in Faculty Evaluation. San Francisco, CA: Jossey-Bass, 1984.
16. Cashin WE. Developing an Effective Faculty Evaluation System. (IDEA Paper No. 33). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development, 1996.
17. Hativa N, Raviv A. Using a single score for summative evaluation by students. Res Higher Educ. 1993;34:625–46.
18. Palmer L. Influence of students' global constructs of teaching effectiveness on summative evaluation. Educ Assess. 1998;5:111–25.
19. Patrick J, Smart RM. An empirical evaluation of teacher effectiveness: the emergence of three critical factors. Assess Eval Higher Educ. 1998;23:165–78.
20. Skeff KM, Stratos GA, Bergen MR, Regula DP. A pilot study of faculty development for basic science teachers. Acad Med. 1998;73:701–4.
21. Richards BF, Wilking AP, Kirkland, RT. A four-month faculty development curriculum on teaching and learning. Acad Med. 1999;74:614–5.
22. Neale V, Roth LM, Schwartz KL. Faculty development using evidence-based medicine as an organizing curricular theme. Acad Med. 1999;74:611.
23. Gjerde CL, Albanese M, Howard N. A faculty development program in basic teaching skills. Acad Med. 1999;74:610–1.
24. Walling AD. Medical education 101—a faculty development course. Acad Med. 1999;74:609.

119

Modeling the Effects of a Test Security Breach on a Large-scale Standardized Patient Examination with a Sample of International Medical Graduates

ANDRÉ F. DE CHAMPLAIN, MARY K. MACMILLAN, MELISSA J. MARGOLIS, DANIEL J. KLASS, ELLEN LEWIS, and SUE AHEARN

Score validity is of central concern to any organization or school involved in high-stakes testing.[1] Validation research entails clearly identifying the purpose for which test scores are to be used so that appropriate empirical evidence can be gathered to substantiate the intended score-based inferences.[2] The validity of these score-based interpretations can be weakened by several test-related phenomena, including breaches to the security of the environment. The impacts of various forms of test security breaches need to be clearly addressed to determine the extent to which a priori knowledge of materials might provide an undue advantage to subgroups of examinees. This evidence also ensures that misinterpretation of scores is minimized on the part of the user. This task is especially crucial with performance-based tests such as standardized patient (SP) examinations, given the typically limited nature of case banks, the long exposure of items/cases, and the high costs associated with developing these types of assessments.[3]

Impact of Security Breaches on Test Performance

The literature devoted to assessing the impacts of various forms of security breaches on the performances of students completing SP tests has reported mixed findings. Most investigations undertaken in this area have been aimed at determining whether mean scores on SP tests vary significantly when cases are administered throughout an extended interval, ranging from as little as several weeks[4] to as much as an academic year.[5] The authors of these studies have reported that mean station or case scores generally remain stable and that the reuse of identical cases, consequently, appears to have only a minimal impact on the scores of students taking the examination at different periods of time throughout the administration cycle.[4,6–8] However, other research suggests that the reuse of identical cases can yield an increase in overall mean score, prompting a suggestion that the number of common cases be kept at a minimum across forms.[5,9,10] Swartz, Colliver, Cohen, and Barrows[11,12] examined whether collusion among students did affect overall SP test scores in a more systematized fashion by encouraging students who took the examination in the early stages of administration to share as much information as possible about the cases with students scheduled to be tested at a later date. The authors found little evidence that information-sharing among students affected performance.

It is important to underscore that those studies restricted their view of a test security breach to various degrees of (presumed) information-sharing among examinees. It can be argued that complicity among students, although a common form of a test-security breach, is probably one of its most benign manifestations. This is especially likely with low- to moderate-stakes SP examinations, where students' motivation to engage in information sharing is low. In a high-stakes context (e.g., in licensure and certification testing), dishonest coaching organizations and examinees might employ a host of illicit means to obtain and disseminate actual test materials. A study undertaken by De Champlain et al.[13] did model the impact of additional, more severe forms of test-security breaches on examinees' performances such as those that would result from students' having access to formal materials prior to taking the examination. The authors reported that disclosing test materials, whether it be directly to a subgroup of examinees or via a dishonest coaching course, led to significant checklist performance gains for a sample of United States medical graduates (USMGs). However, the impact of disclosure on interpersonal skills (IPS) scores was nil. Although informative, it is important to point out that these findings were based on a small and homogeneous sample with respect to examinees' medical education and clinical skill levels. As such, there is a need for this type of research to be replicated with a more varied sample of examinees, to obtain an estimate of disclosure effects that might generalize to a more heterogeneous population of medical students.

The purpose of the present study was to model the impact of disclosing test materials on SP examination scores with a sample of international medical graduates. Furthermore, it is hoped that ensuing findings will provide a practical estimate of expected effect size within the context of this type of security breach and with this population.

Method

Examination. In this investigation, the SP test assessed the clinical (history taking, physical examination, communication) skills and IPS of physicians about to enter supervised practice. SPs are laypeople trained to portray one of a variety of clinical scenarios. Test candidates rotate through these scenarios (or cases) and encounter patients in a setting intended to reflect an ambulatory care clinic. Case-specific checklists are used to assess examinees' clinical skills. These checklists are composed of dichotomously scored items, each of which represents a single action that is expected to be done by the student. A percent-correct score, corresponding to the number of actions done by the student out of the total number of behaviors listed in a given checklist, is computed for all encounters. IPS are assessed with the Patient Perception Questionnaire (PPQ), a case-independent inventory that is composed of six five-point Likert scale items. A percent-correct PPQ score is also computed and reported to each student for all encounters. Both measurement instruments are completed by the SP following each 15-minute encounter with the student. The same ten cases (chosen from the available pool) were administered to all examinees. The cases were selected to reflect the majority of cells contained in the test blueprint with regard to both skill and content domains.

Scoring Procedure. In this examination, two SPs were trained to portray each case. For any given case, the performing SP portrayed the actual clinical scenario with the examinee, whereas the monitoring SP observed the encounter as it proceeded on a video screen in a separate room. Each student's final percent-correct checklist score reflected the consensus reached by the performing and monitoring SPs as to what constituted the appropriate response to each item. Videotape review was instituted to arrive at a consensus if two or more discrepancies per checklist were noted in any given encounter. Of the 9,625 checklist item responses recorded (77 students × 125 checklist items across the ten cases), videotape review was necessary for 202 (2.10%). The PPQ percent-correct score was derived from the performing SP.

Examinees. Seventy-seven international medical graduates (IMGs), recruited from the Los Angeles metropolitan area, partic-

120

ipated in this study and were blinded to its purpose. All examinees were certified by the Educational Commission for Foreign Medical Graduates, i.e., they had successfully passed the following examinations: Step 1 and Step 2 of the United States Medical Licensing Examination and a test of English-language proficiency. The examinees were paid for their participation and randomly assigned to one of two testing conditions: control or security breach (SB). The testing environment for examinees assigned to the control condition ($n = 32$) was representative of a "normal" assessment situation (i.e., participants received routine prior information about the test but no materials from the examination). In the SB condition, we attempted to model a situation in which actual case materials were disclosed. Examinees in the SB condition ($n = 45$) were directly provided with the checklists for five of the ten cases to be seen (referred to as the exposed cases) as well as the PPQ, and were given one to two hours to review these materials prior to completing the test. Information pertaining to the five non-exposed cases was not disclosed to any of the examinees participating in this study. Cases included in the exposed and non-exposed sets were matched with respect to the main areas of this SP test's blueprint.

Analyses. Two separate analyses of covariance (ANCOVAs) were undertaken to compare the performances of the two groups on the five exposed cases. For both models, the condition factor (control or SB) was treated as the independent variable. The mean percent-correct checklist score on the five non-exposed cases was treated as the covariate in the first ANCOVA, while the mean percent-correct checklist score on the five exposed cases was deemed to be the dependent variable (DV). In the second analysis, the mean percent-correct PPQ score on the five non-exposed was deemed to be the covariate, whereas the mean percent-correct PPQ score on the five exposed cases was treated as the DV.

Results

Mean scores and standard errors on the five exposed cases for examinees assigned to each of the two conditions, adjusted for initial differences in ability between groups, were as follows:

For examinees assigned to the control condition,

- the adjusted mean percent-correct checklist score was 54.53 (SE = 1.48), and
- the adjusted mean percent-correct Patient Perception Questionnaire score was 60.87 (SE = 1.18).

For examinees assigned to the security breach condition,

- the adjusted mean percent-correct checklist score was 59.95 (SE = 1.24), and
- the adjusted mean percent-correct Patient Perception Questionnaire score was 67.03 (SE = 0.99).

A significant group main effect was obtained in the first ANCOVA, $F(1,74) = 7.66$, $p = .0071$. For the exposed cases, the SB group (adjusted M = 59.95%) significantly outperformed the control group (adjusted M = 54.53) on the checklist. Similarly, the mean PPQ score for examinees assigned to the SB condition (adjusted M = 67.03%) was significantly higher than the mean estimated for the control group (adjusted M = 60.87%), $F(1,74) = 15.84$, $p = .0002$.

Conclusions

Results obtained in the present study with a sample of international medical graduates mirror those reported in previous research with USMGs.[11] Disclosing checklist items led to significant performance gains for the examinees assigned to the SB condition. The gain noted in this investigation (5.4%), was, however, slightly lower than that obtained with a sample of USMGs. This is probably attributable to the larger number of cases administered in the test form (ten as opposed to six in the past USMG study). Therefore, the challenge posed to the IMGs was slightly more daunting, as they had to sift through ten cases to identify the clinical scenarios for which they possessed disclosed materials and apply this information accordingly. Nonetheless, the gain noted would concretely translate itself into a 4.4-checklist-item disadvantage over five cases (slightly less than one item per case). This advantage might be inconsequential for most USMGs, who typically perform well above the cut-score on this type of examination.[14] However, it could significantly affect decision consistency for IMGs, whose scores tend to cluster in the vicinity of the pass/fail standard in a larger proportion. The control and SB groups also did differ significantly with respect to their mean PPQ scores, a result that was not found with USMGs.[11] Interestingly, the difference between the two groups (6.2%) was actually larger than the one resulting from disclosing checklist items. This could reflect a difference in interaction styles that is culturally based. Disclosing simple indicators of IPS (such as the Likert-scale items found on the PPQ) to SB group examinees yielded a mean score that was similar to that typically encountered with U.S. medical students. It is also worth noting that the type of case that was most susceptible to the effects of disclosure appears to be population-dependent. For U.S. medical students, prior research suggested that cases involving largely mechanical physical examination maneuvers were the easiest to memorize and consequently reflected the highest performance gains for those examinees with prior knowledge of materials. Divulging materials for cases that primarily require communication and IPS in the interaction with the patient proved to be the most beneficial for our sample of IMGs. Again, these findings appear be indicative of differences in the way our sample of IMGs interacted with the SPs. These results suggest that providing a clear description of the examination and its goal to all examinees prior to the administration (in some form of information bulletin, for example) is necessary to ensure a common understanding of expected behavior on the part of students.

In summary, the results presented in this study provide further evidence that the secure handling of test materials is essential for all examinations, whether they be traditional in format or performance-based. Although the security breach modeled in this investigation was severe (half of the test materials were directly exposed to students), steps can nonetheless be undertaken to minimize the likelihood of materials being disclosed. This, in turn, might lessen the impact of a security breach should checklists or other pertinent information fall into the hands of dishonest individuals.

One obvious strategy that should be adopted with all SP tests is to clearly lay out the flow of materials and restrict access solely to concerned staff so that these individuals can be held accountable for receipt and safekeeping of this information. Delivering the measurement instruments via a computer network also seems advisable, given the greater control that the latter medium can afford and the virtual elimination of a "paper trail." The results of our study also point out the need to increase test development efforts to minimize the likelihood of a security breach. Increasing the pool of available cases enables a more frequent rotation of forms within and across test sites, thus limiting the exposure rate for any given set. Finally, the use of modeled or cloned cases also seems desirable to increase the size of the case pool and thwart those individuals who may have mechanically memorized cases and accompanying materials. Modeled cases are defined as those presenting a similar opening scenario but requiring a different work-up on the part of the student. Cloned cases, on the other hand, call for a similar set of actions on the part of the student but present different contexts.

Although informative, our results need to be interpreted in light of several limitations. First, the sample size examined was small, and generalizations should be made with caution. Our sample was also

121

composed of IMGs who were perhaps atypical of the corresponding population, given that they had successfully fulfilled several U.S. medical licensing requirements (passed the USMLE Step 1 and Step 2 and a test of English-language proficiency). Consequently, the effect sizes reported in this study should probably be viewed as lower-bound estimates of what to expect in an operational testing context. Replication of this research with different groups of both IMGs and USMGs seems advisable. This research might also permit us to test the hypothesis that lower-ability students might benefit more from gaining access to materials than would those who are more proficient. From a test-development perspective, pursuing research that focuses on the identification of characteristics that make a case more vulnerable to memorization would also be helpful. Finally, the findings reported in this study underscore the need to develop methods to detect breaches to the security of the testing environment. Research aimed at assessing the usefulness of "tagged" checklist items and other means should be pursued.[15]

Testing organizations and medical schools should always be vigilant in guarding themselves against dishonest examinees and organizations that may wish to compromise the secure nature of the testing environment. This investigation confirms past findings in that the psychometric properties of the SP examination described appear to be vulnerable to blatant disclosure of testing materials. It is hoped that the results presented in this article will foster future relevant research that will ultimately lead to the implementation of secure SP tests for licensure and other purposes.

Correspondence: André F. De Champlain, PhD, Senior Psychometrician, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104.

References

1. Kane M. The validity of licensure examination. Am Psychologist. 1982;37:911–8.
2. Messick S. Standards of validity and the validity of standards in performance assessment. Educational Measurement Issues and Practice. 1995;14:5–8.
3. Mehrens WA, Phillips SE, Schram M. Survey of test security practices. Educational Measurement: Issues and Practice. 1993;12:5–19.
4. Colliver JA, Barrows HS, Vu NV, Verhulst SJ, Mast TA, Travis TA. Test security in examinations that use standardized-patient cases at one medical school. Acad Med. 1991;66:279–82.
5. Cohen R, Rothman AI, Ross J, Poldre P. Impact of repeated use of objective structured clinical examination stations. Acad Med. 1993;68(10 suppl): S73–S75.
6. Colliver JA, Travis TA, Robbs RS, Barnhart AJ, Shirar LE, Vu NV. Test security in standardized-patient examinations: analysis with scores on working diagnosis and final diagnosis. Acad Med. 1992;67(10 suppl):S7–S9.
7. Niehaus AH, DaRosa DA, Markwell SJ, Folse R. Is test security a concern when OSCE stations are repeated across clerkship rotations? Acad Med. 1996;71:287–9.
8. Stillman PL, Haley HA, Sutnick AL, et al. Is test security an issue in a multistation clinical assessment? A preliminary study. Acad Med. 1991:66(10 suppl): S25–S27.
9. Jolly BC, Newble DI, Chinner T. The learning effect of re-using stations in an objective structured clinical examination. Teach Learn Med. 1993;5:66–71.
10. Jolly BC, Cohen R, Newble DI, Rothman AI. Possible effects of reusing OSCE stations. Acad Med. 1996;71:1023–4.
11. Swartz MH, Colliver JA, Cohen DS, Barrows HS. The effect of deliberate, excessive violations of test security on performance on a standardized-patient examination. Acad Med. 1993;68:76–8.
12. Swartz MH, Colliver JA, Cohen DS, Barrows HS. The effect of deliberate, excessive violations of test security on a standardized-patient examination: an extended analysis. In. Rothman AI, Cohen R (eds). Proceedings of the Sixth Ottawa Conference on Medical Education. Toronto, Ontario, Canada: University of Toronto Bookstore Publishers, 1994:280–4.
13. De Champlain AF, Macmillan MK, Margolis MJ, et al. Modeling the effects of security breaches on students' performances on a large-scale standardized patient examination. Acad Med. 1999;74(10 suppl):S49–S51.
14. Margolis MJ, De Champlain AF, Klass DJ. Setting examination-level standards for a performance-based assessment of physicians' clinical skills. Acad Med. 1998; 73(10 suppl):S114–S116.
15. Macmillan MK, De Champlain AF, Klass DJ. Using tagged items to detect threats to security in a nationally administered standardized patient examination. Acad Med. 1999;74(10 suppl):S55–S57.

122

ISI

Assessing Post-encounter Note Documentation by Examinees in a Field Test of a Nationally Administered Standardized Patient Test

MARY K. MACMILLAN, ELIZABETH A. FLETCHER, ANDRÉ F. DE CHAMPLAIN, and DANIEL J. KLASS

The large-scale standardized patient (SP) test in this study assessed the clinical skills of fourth-year medical students in a series of clinical encounters targeting history taking, physical examination, communication, and interpersonal skills. Yearly large-scale field tests have been undertaken over the past seven years in preparation for national administration. The study reported here was conducted in 1998.

Students are oriented to the test prior to completing up to 12 15-minute SP encounters (cases). Following each encounter, the SP records history elicited, counseling provided, or physical examination performed using an objective checklist developed by expert clinicians. The checklists may be thought of as a process measure, serving as a reflection of actual behaviors demonstrated by the candidate. Interpersonal skills are assessed using the Patient Perception Questionnaire (PPQ), a six-item instrument with a five-point Likert rating scale (uniform for every case).

Following each encounter, students are given seven minutes to write a free-response Post-Encounter Note (PEN) (either a list of significant positive and negative history and physical findings or a written chart note documenting findings and counseling). The PEN is specifically tailored to reflect each case. There is no limit to the number of findings students may write. Patient management (diagnosis or therapeutic plans) and interpretation of diagnostic tests are not assessed in these PENs. The PENs potentially reflect a candidate's ability to determine the most significant findings elicited from the encounter and to accurately record them. While numerous studies have examined the use of checklists with respect to fairness, security, and accuracy, there is limited research investigating the psychometric properties of PENs.

Previous studies have examined appropriate methods for scoring the PEN. Soliciting global judgments from experts seems appealing because scores are derived from the expertise of practicing physicians,[1] but global ratings can be unreliable unless the scoring task is highly structured and extensive standardized training is provided.[2,3] From a national testing perspective, recruiting physicians to score the PENs for thousands of candidates may not be feasible. As a result, many researchers have favored the use of analytic keys to score PENs.[2,4] A significant advantage of using such scoring keys is the fact that non-physicians can be trained to score the PENs with an accuracy level comparable to that of physicians.[5-7]

Research examining the usefulness of PENs with an SP test has suggested that these scores contribute valuable information to the assessment of clinical skills by providing unique information different from that derived from checklist scores.[8] However, other research indicates that the chart audit scores should not replace the checklist entirely, since the information written by candidates in a simulated medical record may not provide a complete picture of events during an SP encounter.[9]

The inclusion of the PEN in an SP test is appealing. First, it is thought that PENs are relatively immune to within-site and cross-site effects. Also, they do not depend on the accurate recording of checklists by SPs. Additionally, threats to security are minimized because the PEN is a free-response instrument and does not reveal checklist content or other exam material.

However, before the PEN can be used in large-scale testing, it is important to determine whether the PEN is a reflection of the checklist or whether the PEN contributes unique information about a student's ability to synthesize and record medical information. The purpose of this study was, therefore, to investigate the relationship between entries recorded in the PEN and actions captured on the checklist. It is hoped that the results of this study will help determine how to best incorporate PEN information into a composite score.

Methods

Measurement Instruments. History taking, physical examination, and communication skills in this SP test were assessed using six case-specific checklists composed of dichotomously scored items. Each checklist focused on two of the three clinical skills and contained a maximum of 25 items deemed critical for that case by a panel of expert clinicians. The six cases used in the study were based on a test blueprint reflecting the challenges expected to be encountered by a medical student entering residency. Checklists and PPQs were completed by the SPs following each 15-minute clinical encounter.

An SP monitor viewed the encounter on a video screen and completed the same checklist as the encounter unfolded. If two or more item responses on the checklist completed by the SP conflicted with the responses on the checklist provided by the SP observer, a review of the videotaped encounter was required. The SP and the SP monitor reviewed the videotape together and discussed what constituted a correct response. The agreed-upon response became the key.

A panel of expert clinicians was selected from a range of medical specialties. Each member of the panel had experience teaching third- or fourth-year medical students and experience with developing or implementing SP exams for clerkships or other courses. The panel developed analytic scoring keys for the PENs of the six study cases by using checklists, guides to the checklists, case prescriptions, and videotapes of each clinical encounter. These keys contained a list of significant findings or acceptable synonyms that the clinicians deemed essential for inclusion in a patient note.

Examinees. The sample was composed of 80 fourth-year medical students from one northeastern medical school. The students were required to participate in the examination, but the scores did not contribute to their end-of-year grades.

A group of six medical chart abstractors (MCAs) was recruited to score the PENs of the 80 students. The MCAs were divided into three pairs and each pair was assigned to score the PENs for two cases. The MCAs were oriented to each of their assigned cases by listening to the case summary and student instructions and by watching a videotaped encounter of the case. The MCAs then scored an initial sample of PENs for each case using the keys developed by the panel. The MCAs were instructed to give credit for each finding (or acceptable synonym) included in the PEN. Following this "practice scoring exercise," the pairs discussed their decisions and reached consensus. The MCAs then scored the remaining PENs and completed a post-scoring survey regarding their reactions to the process.

Analyses. Items on the scoring key were matched to content-equivalent items on the checklist by a test-development staff professional. Items on the scoring key without a matching checklist item and exemplars were not reviewed in the study. For each

matched scoring key and checklist item, the proportion-of-agreement rates (the observed agreement rate between a given behavior as measured by the checklist and the PEN) as well as kappa coefficients (an estimate of agreement above and beyond that expected due to chance) were computed.

Results

The frequency distributions of the proportion-of-agreement rates and kappa coefficients for the six cases are shown in Table 1. A total of 68 item comparisons were examined across six cases. Overall, 69% of the items had proportion-of-agreement rates that fell between .61 and 1.0. At the case level, 75% of the items in Case 1, 92% of the items in Case 2, 69% of the items in Case 3, 56% of the items in Case 4, 100% of the items in Case 5, and 36% of the items in Case 6 had proportion-of-agreements rates that fell between .61 and 1.0. Across the six cases, 31% of the items had rates that fell between .21 and .60, and none of the items had a proportion-of-agreement rate below .20. Kappa values ranged from −.12 (Case 6) to .90 (Case 3). Kappa cannot be computed for items that are correctly answered by all examinees, hence, the total number of item comparisons across all six cases was 62. Using classification guidelines proposed by Landis and Koch,[10] perfect or excellent agreement was achieved for 15% of the items, good or fair agreement for 42%, and slight or poor agreement for 44%.

The average proportions of discordance between items on the checklists and the PENs are provided in Table 2. Column three in the table represents the average proportion of students who received credit for an item during the encounter (checklist = Y, or yes) but did not write the findings in the note (PEN = N, or no, or errors of omission). At the skill level, the average proportions ranged from .08 to .31 for history taking, .09 to .33 for physical examination, and .09 to .39 for communication skills. Poor documentation in the note does not appear to have been skill-specific. Column five shows the average proportions of students who recorded findings in the note (PEN = Y, or errors of commission) that were not actually pursued in the encounter (checklist = N). With the exception of that for the history-taking skill in Case 6, these average proportions are considerably less than those in column three.

Discussion

Proportion-of-agreement rates uncorrected for chance were strong for four of the six cases. Cases 4 and 6 exhibited moderate agreement. The proportion-of-agreement rates corrected for chance were poor to moderate. Cases 4 and 6, for which agreement was particularly poor, required that the student use appropriate patient education and counseling techniques and reassure the patient. Thus, students appeared to have difficulty synthesizing and appropriately recording the psychosocial elements of an encounter. It may be that the students simply did not recognize the pertinence of documenting information related to patient education and counseling. Other studies have also reported poor documentation of items related to patient education.[5]

On the whole, however, poor documentation in the PEN was not confined to cases with communication components. Omitting from the PEN items that were actually pursued during the encounter appeared to some degree in every case. In particular, the physical examination items in Cases 1 and 3 were not documented well, suggesting that the students had difficulty interpreting physical exam findings into the written word. Again, it may be that the students simply did not understand the significance of the information gleaned from the encounter. It is also possible that the students' concept of an adequate note was less comprehensive than that envisioned by the test developers. Revising the student orientation to the test with an emphasis on the importance of the PEN and providing examples of appropriate documentation for a clinical encounter may improve documentation rates.

Although it was less likely that the students recorded in the PEN findings that were not actually pursued during the encounter, such items occurred more often in Cases 2 and 6 for history taking, and in Case 4 for communication items. At first glance it appeared that the students might have fabricated findings, but closer examination of the items where this occurred revealed potential problems with the wording of the scoring key. For example, while a checklist item is very specific, a single PEN item attempts to capture all possible synonyms; thus, overlapping concepts may be inadvertently lumped together. Identifying the discrepancies between the checklist and the PEN may be helpful for refining the scoring key.

It is important to interpret our findings in light of limitations

Table 1. Frequency Distribution of Proportions-of-agreement Rates (and Kappa Coefficients) between Checklist and Post-Encounter Note (PEN) Items in a Field Test of an Administered Standardized Patient Test, 1998*

	Proportion of Agreement Rates			Kappa Coefficients		
	0 to .20	.21 to .60	.61 to 1.0	0 to .20	.21 to .60	.61 to 1.0
Case 1	0	2	6	2	4	0
Case 2	0	1	12	2	7	3
Case 3	0	4	9	4	5	3
Case 4	0	7	9	10	4	2
Case 5	0	0	7	1	3	1
Case 6	0	7	4	8	3	0

*Eighty fourth-year medical students at a northeastern medical school took the SP examination of six stations (15 minutes each). The examination was required but the score did not contribute to the end-of-year grade.

Table 2. Average Proportions of Checklist–Post-Encounter Note (PEN) Discordance, by Case and Skill, of Examinees' Performances on a Field Test of a Nationally Administered Standardized Patient Test, 1998*

Skill	No. of Item Comparisons	Average Proportion: Checklist = Yes, PEN = No	Average Proportion: Checklist = No, PEN = Yes
Case 1			
History	3	.22	.00
Physical exam	5	.32	.03
Case 2			
History	10	.16	.08
Physical exam	3	.09	.01
Case 3			
History	5	.08	.01
Physical exam	8	.33	.05
Case 4			
History	8	.31	.03
Communication	8	.23	.12
Case 5			
History	5	.15	.01
Communication	2	.09	.01
Case 6			
History	7	.19	.23
Communication	4	.39	.01

*Eighty fourth-year medical students at a northeastern medical school took the SP examination of six stations (15 minutes each). The examination was required but the score did not contribute to the end-of-year grade.

LSI - 124

inherent in the study. First, the analyses were conducted with a small sample (80 students per case) from a single medical school; thus, generalizations should be tentative. Second, the study examined only six cases, which provided a limited opportunity to sample across the test blueprint. Future research should address these limitations by expanding the focus of the study to several diverse schools. Increasing the number of cases and the sample size will also help to minimize measurement error. Finally, excluding items from the scoring key without a matching checklist item may have underestimated the students' abilities to document their findings. Within this study, the PEN scoring key was assumed to represent the "gold standard" for appropriate documentation. Students may have written in the notes items that would not be credited because the clinical experts did not deem them critical. Further content validation of the scoring key may be useful in addressing this shortcoming.

Despite these limitations, the results of this study suggest that the PEN provides unique information about students' abilities to document the gathering of information, understand the significance of the information gathered, and translate verbal information into the written word. Thus, poor concordance between the checklist and the PEN suggests that students may have limited skills to properly synthesize, interpret, and record findings from a clinical encounter. Given these results, test developers may be less inclined to treat the PEN score as simply a reflection of the checklist or redundant information, in view of the fact that the ability to record significant findings from clinical encounters and demonstrate understanding of those findings is a critical function not only of medical students but also of practicing physicians.

References

1. Boulet J, Friedman Ben-David M, Ziv A, Burdick WP, Gary NE. The use of holistic scoring for post-encounter written exercises. In: Proceedings of the Eighth International Ottawa Conference on Medical Education and Assessment, 1998 July 12–15, Philadelphia, PA. Philadelphia, PA: National Board of Medical Examiners, 1998.
2. Frijns PHAM, Van Der Vleuten CPM, Verwijnen GM, Van Leeuwen YD, Wijnen WHFW. The effect of structure in scoring methods on the reproducibility of scores of tests using open-ended questions. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP (eds). Teaching and Assessing Clinical Competence. Groningen, The Netherlands: Boek Werk Publications, 1990;466–71.
3. Van Der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ. 1996;1:41–67.
4. Finkbiner R, Fletcher E, Orr N, Klass DJ. Question format and scoring methods for standardized patient interstation exercises. In: Rothman AI, Cohen R (eds). Proceedings from the Sixth Ottawa Conference on Medical Education. Toronto, ON, Canada: University of Toronto Bookstore, 1994;343–5.
5. De Champlain A, Fletcher E, Macmillan M, Klass D, Margolis M. Assessing the reliability of post-encounter note scores in a large-scale standardized patient examination: comparing the scoring consistency of medical chart abstractors and physicians. In: Proceedings of the Eighth International Ottawa Conference on Medical Education and Assessment, 1998 July 12–15, Philadelphia, PA. Philadelphia, PA: National Board of Medical Examiners, 1998.
6. Friedman Ben-David M, Boulet JR, Burdick WP, Ziv A, Hambleton RK, Gary NE. Issues of validity and reliability concerning who scores the post-encounter patient-progress note. Acad Med. 1997;72(10 suppl):S79–S81.
7. Stillman PL, Regan MB, Haley HA, Norcini JJ, Friedman M, Sutnick AI. The use of a patient note to evaluate clinical skills of first-year residents who are graduates of foreign medical schools. Acad Med. 1992;67(10 suppl):S57–S59.
8. Colliver JA, Travis TA, Robbs RS, Vu NV, Marcy ML, Barrows HS. Assessment of uniqueness of information provided by postencounter written scores on standardized-patient examinations. Eval Health Prof. 1992;15:465–74.
9. Case SM, Templeton B, Samph T, Best AM. Comparison of observation-based and chart-based scores derived from standardized patient encounters. In: Harden R, Hart I, Mulholland H (eds). Approaches to Assessment of Clinical Competence. Norwich, England: Page Brothers, 1992:471–5.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–73.

International Medical Graduates' Performances of Techniques of Physical Examination, with a Comparison of U.S. Citizens and Non–U.S. Citizens

STEVEN J. PEITZMAN, DANETTE MCKINLEY, MICHAEL CURTIS, WILLIAM BURDICK, and GERALD WHELAN

Literature dating back over 25 years has documented and commented upon deficiencies in the performances of medical students and house officers in both techniques of physical examination and ability to detect abnormalities.[1–7] Many of these studies took place in U.S. teaching hospitals, and when the subjects were residents, the studies do not specify results for international medical graduates (IMGs). Yet in 1998–1999 25% of first-year house officers in U.S. postgraduate medical training programs were IMGs,[8] whose clinical experience in medical school is considered more variable than that offered in U.S. and Canadian schools.[9] A substantial number of IMGs are actually American and Canadian citizens acquiring their undergraduate medical training outside North America, particularly at the "offshore" schools, which have proliferated in the Caribbean. The quality of training in clinical skills in these new schools, which are not accredited by the Liaison Council on Medical Education, is largely unknown. In July 1998 the Educational Commission for Foreign Medical Graduates (ECFMG) implemented its Clinical Skills Assessment (CSA®), a new requirement for ECFMG certification. Experience with a more recently created "physical examination case" within the CSA has allowed us to measure skills in a selection of basic physical examination techniques among IMGs completing this high-stakes performance assessment, including non-U.S. citizens and U.S. citizens. Both groups were deficient in important skills.

Methods

Test Case and Design. The CSA is a ten-station performance assessment using standardized patients (SPs). It is designed to measure capabilities in history taking, certain aspects of physical examination, oral and written communications, interpersonal behaviors, and the English language. The typical case requires the examinee to assess a new patient problem by taking a focused history and performing what the examinee considers a relevant physical examination. The examinee completes a "patient note" and suggests a differential diagnosis and a diagnostic plan. The SPs use checklists to document which expected elements of history taking and physical diagnosis the candidate did, but the format cannot always distinguish whether an examinee omitted a physical examination element or attempted it but made an error. For this reason, the physician staff of the CSA, with the endorsement of its Test Development Committee, created a "physical examination case." The case scenario presents a young man who needs a pre-employment physical examination; the "patient" hands to the examinee a simulated examination form that explicitly indicates the physical examination components to be done (listed in Table 1). These elements were chosen not to replicate an entirely realistic pre-employment examination, but rather to include tasks relating to a variety of organ systems and to include some for which correct technique would likely be especially necessary for detecting abnormalities in actual practice.

Each task, such as "auscultation of lungs," "ophthalmoscopic examination," "deep tendon reflexes," was broken down into one to four components or scoring criteria, each scored by the SP as "done" or "not done" by the examinee. For example, for ophthalmoscopic examination, a candidate can separately obtain a point for correctly instructing the patient, for using his or her right eye for patient's right eye and left for left, and for bringing the instrument sufficiently close to the patient's eye. Such criteria were largely based on techniques outlined in a standard textbook on physical examination that is used extensively both in the United States and in other countries.[10] Criteria for auscultation of the heart were minimal because we could not fairly expect special positioning and maneuvers in the setting of a screening examination for a young adult without complaints. More criteria might have been entered for some tasks, but since the CSA intends to assess basic clinical skills, we aimed at the most essential techniques. Also, we did not want to impair SP recall with too long a checklist.

Table 1. Performances of Non–U.S.-Citizen and U.S.-Citizen International Medical Graduates (IMGs) of Selected Physical Examination Techniques Tested in One Case within the Clinical Skills Assessment (CSA®) of the Educational Commission for Foreign Medical Graduates, October 1999–January 2000

		Mean Score (95% Confidence Limits)		
Physical Examination Task	No. of Scoring Criteria*	(n = 318)* (%)	Non-U.S. IMGs (n = 247)* (%)	U.S. IMGs (n = 71)* (%)
Measure blood pressure	3	87 (83–92)	84 (80–88)	91 (83–98)
Assess extraocular movements	1	83 (78–89)	75 (70–80)	92 (82–100)†
Ophthalmoscopic examination	3	70 (65–74)	60 (55–64)	80 (72–88)†
Percussion of lungs	2	76 (71–81)	76 (71–80)	76 (67–85)
Auscultation of lungs	3	91 (88–94)	90 (87–92)	93 (88–98)
Auscultation of heart	1	99 (97–100)	99 (97–100)	99 (96–100)
Radial and dorsalis pedis pulses	2	80 (75–85)	72 (67–77)	88 (79–96)†
Deep tendon reflexes	4	80 (76–84)	74 (70–78)	86 (78–93)†
Whole case	19	79 (77–81)	77 (74–79)	87 (83–90)†

*"Scoring criteria" are components of the physical examination task. Mean scores were averages over the groups of the proportions of criteria correctly met by examinees.
†Difference in means between non-U.S. and U.S. IMGS is significant, $p < 0.01$.

126

The case was carried out by one SP following intensive training. One author (SJP) validated the SP's accuracy using simultaneous checklist scoring during "pilot" runs of the case. The SP was already considered by staff a rapid learner and accurate in his work, and showed good scoring concurrence during quality-assurance observations in this and another case (less than 10% discrepancy). He has no physical abnormalities.

The case is not used in every administration of the test. It is one of a group of "miscellaneous" cases chosen by a computerized selection program designed to achieve balanced forms while accommodating the availability of SPs. From the perspective of any one candidate, the appearance of this case on her or his ten-case form was effectively random: the candidates were in no way prospectively selected. We report on the first 318 candidates who encountered this case on their form—247 non-U.S. citizens and 71 U.S. citizens from October 1999 through January 2000. The ratio of U.S. to non-U.S. citizens (.29) in this group turned out to be somewhat lower than the ratio (.44) for all 8,313 candidates tested by the CSA as of the date of analysis. The overall test scores in data gathering (history taking and physical examination) across all ten cases in their forms for the 318 examinees in this study were similar to those for all candidates ($t_{1,8612}$ = 1,732, p = 0.08), suggesting that our cohort was representative.

Analysis. For a task, such as "deep tendon reflexes," comprising four scoring criteria, an examinee could obtain 0, 1, 2, 3, or 4 points, expressed for each task as a percent-correct score of 0%, 25%, 50%, 75%, or 100%, with a similar transformation used for tasks comprising fewer subtasks. We calculated the mean of these percent-correct scores for each task (e.g., deep tendon reflexes), and for the whole case (percentage of all 19 criteria done correctly), over all examinees in the cohort, and did the same for the subgroups of U.S. IMGs and non-U.S. IMGs. Confidence intervals were also calculated. To compare the performances of non-U.S. IMGs and U.S. IMGs, we conducted a repeated-measures analysis of variance. The eight physical examination tasks were the within-subject factors, and citizenship at start of medical school was the between-subjects factor. Post-hoc analyses were conducted to determine whether differences in task scores between groups were significant.

To better understand qualitatively the nature of frequently scored errors and omissions, the author (SJP) most responsible for designing the case and training the SP interviewed the SP and observed 40 randomly selected tapes of actual encounters. The SP was asked, where appropriate, to recall the most common errors causing him to withhold a mark for a given scoring criterion (e.g., "palpating too high on the foot" for dorsalis pedis pulse).

Results

Table 1 shows the mean percentage scores for each task and for the whole case for all examinees in the cohort and for the U.S. and non-U.S. subgroups. The task main effect was significant (F = 22.631, $p < .01$), indicating that the tasks, averaged over the two groups, were not of equal difficulty. The weakest performance was in ophthalmoscopy, the strongest in cardiac examination (for which, as mentioned, the criteria were minimal). There was a significant group (between-subjects) effect (F = 14.325, $p < .01$), indicating that, averaged over the eight tasks (or the whole case), there was a statistically significant difference in scores between the two groups. The U.S. IMGs obtained significantly higher case scores than did the non-U.S. IMGs.

The group-by-task interaction was also significant (F = 4.126, $p < .01$), indicating that differences in performances between groups varied over the eight tasks. That is, the U.S. IMGs performed significantly better than did the non-U.S. IMGs for extraocular movements, ophthalmoscopic examination, locating radial and dorsalis pedis pulses, and deep tendon reflexes.

Analysis of the scores for each scoring criterion within the eight physical examination tasks (not presented here), a "debriefing" interview with the SP, and review of a sample of videotapes of encounters revealed the following common technical deficiencies: clumsiness in properly placing and wrapping the blood pressure cuff; insufficient extent of induction of motion in testing eye movement; failure to use "right eye for right eye and left eye for left eye" and not bringing the instrument in closely enough for ophthalmoscopic examination; not comparing right with left at a given location on the thorax for pulmonary percussion; unfamiliarity with the location of the dorsalis pedis pulse; lack of briskness in applying the reflex hammer and applications at incorrect locations.

Discussion

Our study looked only at proficiencies in some fundamental techniques of physical examination, not the ability to recognize and interpret abnormalities. Thus, performance levels below 90% of criteria met can be considered a cause for some alarm when observed in medical school graduates, or final-year students, intending to enter a postgraduate training program.[6] While few prescribed and traditional techniques in physical diagnosis have been rigorously tested to determine whether they improve accuracy in detecting or excluding abnormalities, we used as criteria well-established methods advocated in the most widely used textbook of physical diagnosis. Furthermore, it is difficult to deny that little will be seen in the fundus by an examiner holding the instrument 10 inches from the eye, or that a meaningful interpretation of the deep tendon reflexes is unlikely to follow misapplication of the hammer. It is not too much to expect that every new house officer on the first day of residency would be able to effortlessly and rapidly apply and use the sphygmomanometer in an urgent situation, yet our cohort of IMGs showed only an 87% level of proficiency in this skill. Of interest, McKay et al. tested Canadian medical graduates and found deficiencies in the technique of blood pressure measurement, though they used a more stringent set of criteria than ours.[11]

The ophthalmoscopic examination warrants comment. Non-U.S. IMGs showed only a 60% and U.S. IMGs an 80% level of proficiency, a significant difference but low score for both. Recent literature[5,7] and the observations of one of us (SJP) at the medical school where he teaches suggest a declining use of the ophthalmoscope among learners and teachers in American academic medicine. Our results in this study hint that the situation is similar elsewhere. McNaught and Pearson, in the United Kingdom, found that ownership of an ophthalmoscope declined sharply after an "equipment grant" was discontinued.[12] While any conception of the core skills in physical diagnosis must evolve to match changing patterns of practice,[7] arguably all general physicians and some non-ophthalmologic specialists should be able to recognize at least papilledema, the advanced optic cupping of glaucoma, and perhaps some of the findings associated with common diseases such as diabetes and hypertension. Faulty basic technique, as evidenced by the IMGs we tested, will both frustrate those trying to master this difficult element of physical diagnosis and impede accuracy.

Why might non-U.S. IMGs have performed less well in some tasks than U.S. IMGs? The ECFMG elected to create and implement the CSA based in part on the belief that clinical instruction among international medical schools is less standardized and more variable in extent than that offered by U.S. and Canadian schools accredited by the Liaison Committee for Medical Education.[9] A majority of U.S. IMGs taking the CSA have attended one of the "offshore" medical schools. Students in these schools do much of their third- and fourth-year clinical rotations in U.S. hospitals and practices, and so may encounter the sorts of physical diagnosis expectations tested for in the CSA. Candidates have the opportunity to try out the physical examination equipment available in our examination rooms before the examination begins. Staff have on

127

several occasions heard non-U.S. IMGs report that they had never used an ophthalmoscope or (more rarely) had seldom performed a blood pressure measurement. We are not aware, however, of any comparison of preliminary clinical skills instruction among U.S./Canadian, "offshore," and other international medical schools.

We do not believe that the SP performing this case showed bias in favor of U.S. IMGs over non-U.S. candidates. Obviously our training program for SPs includes discussion of bias and the imperative to avoid it. Also, by chance, the SP chosen for this case is himself a native of another country and speaks with an accent. Furthermore, our observations of a sample of encounters seemed to confirm the differences detected.

Our study has limitations. It provides no comparison of skills of IMGs with those of graduates of U.S. and Canadian schools, and we by no means intend to imply that the latter would not show some deficiencies—indeed, literature cited earlier suggests that they would. We were not able to assess all commonly used physical examination tasks, and such skills as rectal, pelvic, and breast examination are not incorporated into the CSA. As noted, our physical examination case yields little information on ability to carry out a thorough cardiac examination appropriate to a symptomatic patient. Observations of videotapes revealed that occasional candidates did not attend to the explicit instructions for the case and failed to attempt one or more tasks, though we do not think the resulting invalid scores would influence the overall results and conclusions.

This study has several implications. First, residency program directors should be aware that some medical graduates entering their programs might not bring with them a full repertoire of fundamental skills in physical examination technique; of course, our results apply only to graduates of medical schools outside the United States and Canada. It therefore may be desirable to assess selected clinical skills early in the first year and provide focused remediation for detected errors. Second, those responsible for clinical skills instruction at the medical school level may need to also sharpen their focus on ensuring the acquisition of fundamental physical diagnosis methods before students graduate. Finally, the authors' experience with this station supports the now widely accepted view that well-trained standardized patients can be used to assess ability in at least rudimentary techniques of physical examination.

Correspondence and requests for reprints: Steven Peitzman, MD, ECFMG, 3624 Market Street, Philadelphia, PA 19104; e-mail (speitzman@ecfmg.org).

References

1. Wiener S, Nathanson M. Physical examination: frequently observed errors. JAMA. 1976;236:852–5.
2. Wray N, Friedland J. Detection and correction of house staff errors in physical diagnosis. JAMA. 1983;249:1035–7.
3. Johnson J, Carpenter J. Medical house staff performance in physical examination. Arch Intern Med. 1986;146:937–41.
4. Fred H. Requiem for the ophthalmoscope. Hosp Pract. 1994 Feb;29:37–8.
5. Li J. Assessment of basic physical examination skills of internal medicine residents. Acad Med. 1994;69:296–9.
6. Mangione S, Peitzman S. Physical diagnosis in the 1990s: art or artifact? J Gen Intern Med. 1996;11:490–3.
7. Mangione S, Peitzman S. Revisiting physical diagnosis during the medical residency: it is time for a logbook—and more. Acad Med. 1999;74:467–9.
8. Miller R, Dunn M, Richter T. Graduate medical education, 1998–1999: a closer look. JAMA. 1999;282:855–60.
9. Ben-David M, Klass D, Boulet, J, et al. The performance of foreign medical graduates on the National Board of Medical Examiners (NBME) standardized patient examination prototype: a collaborative study of the NBME and the Educational Commission for Foreign Medical Graduates (ECFMG). Med Educ. 1999;33:439–46.
10. Bates B, Bickley L, Hoekelman R. A Guide to Physical Examination and History Taking. 6th ed. Philadelphia, PA: J. B. Lippincott, 1995.
11. McKay D, Raju M, Campbell N. Assessment of blood pressure measuring techniques. Med Educ. 1992;26:208–12.
12. McNaught A, Pearson R. Ownership of direct ophthalmoscopes by medical students. Med Educ. 1992;26:48–50.

● PREPARING TO MAKE THE GRADE *Moderator: Lynn Epstein, MD*

Comparison of Three Parallel, Basic Science Pathways in the Same Medical College

DAVID P. WAY, ANDY HUDSON, and BRUCE BIAGI

Since 1970, the Ohio State University College of Medicine and Public Health has offered medical students a choice between two basic science pathways, lecture discussion (LD) and independent study (IS). Since 1991 the college has offered entering students a choice among three pathways, LD, IS and problem-based learning (PBL). Most of the literature on implementing alternative basic science curricula has focused on the comparison of USMLE Step 1 test scores between different curricular methods. The purpose of this study was to investigate outcome measures (other than USMLE test scores) such as student activities and achievement in clinical education, and affective measures of student and faculty satisfaction. Additionally, we sought to assess the effect of pathway choice on admission, and to determine the factors influential in determining student pathway choice.

Ours is the only medical school in the country where entering students have a choice of three preclinical pathways, making it fertile ground for comparison of the effects of different curricula. Learning objectives, content material, and structure (organ-based organization) are very similar across all three pathways. The three also share faculty, staff, and administrative oversight. What differs across pathways are the teaching and learning methods.

In 1997–98 the college formed a task force to study the benefits and overall desirability of maintaining the three preclinical pathways. Specifically, the task force was charged to look at all three pathways in terms of their educational importance, student and faculty preferences, and participant satisfaction.

Until recently, the traditional LD was the most commonly chosen pathway among the 210 matriculating students each year. The primary mode of teaching in this pathway is large-group lecture supplemented with small-group discussions and labs. The IS pathway, established in 1970 as the first alternative to the LD, offers students the flexibility to learn on their own through the use of highly structured reading materials, computer-based materials, and diagnostic practice examinations. The PBL pathway, established in 1991, emphasizes student-centered, self-directed learning. Unlike IS students, PBL students are introduced to basic science concepts through the analysis and discussion of clinical cases during small-group meetings. Students then work independently on learning issues that are defined by the group before coming back together to discuss their studies.

Literature Review

Like any educational innovation, both IS and PBL programs have had to prove their effectiveness as alternatives to the traditional lecture-based teaching. Lecture-based teaching has existed primarily for its efficiency, not necessarily for its effectiveness.

As medical schools struggled to develop alternatives to lectures, investigations comparing alternatives to traditional lecture curricula such as IS and PBL were reported in the literature. Such investigations have generally found little or no difference in examination scores or clinical performances when comparing lecture-based courses with alternatives. Way et al. compared alternative curricular approaches in one college and confirmed that no difference in average USMLE Step 1 scores existed across alternative basic science pathways when controlling for pre-matticulation differences.[1]

The literature on IS in the health professions reveals the following:

1. There is little or no significant difference in learner performances as measured by examinations and patient care compared with traditional lecture-based curricula.[2–11]
2. IS offers both faculty and students more flexibility and portability in learning when compared with lecture-based learning.[2,3,8]
3. IS promotes lifelong, independent learning, self-pacing, and self-responsibility in learning.[2,11]
4. Students who participate in IS tend to pursue more research and full-time faculty positions than students in lecture programs.[2]
5. After start-up costs are accounted for, IS costs the same as or less than traditional lecture-based courses.[2,3,7]

The literature on PBL in the health professions reveals the following:

1. There is little or no significant difference in learner performances as measured by examinations or patient care compared with traditional lecture-based curricula.[12,16,17]
2. Differences that have been reported generally indicate the same or less factual knowledge but better clinical performance and patient management for PBL students.[13–15,19]
3. Both faculty and students find PBL more enjoyable and prefer PBL to "traditional" lecture courses.[12–14,16,18,19]
4. PBL students tend to use "backward" reasoning (working from clinical information back to theory) when solving clinical problems, whereas traditional students reason "forward" (from theory to clinical practice).[13,15,21]
5. PBL students have a greater tendency to use evidence-based medicine practices (more journals and literature searches) than "traditional" students.[14,15]

Method

This article reports part of a larger, more comprehensive institutional research project conducted by a task force of clinical and basic science faculty supported by consultants from the College of Medicine's Office of Academic Services (OAS) for Medical Education. Both qualitative and quantitative data were gathered for this report using a variety of methods: document analysis, survey methods, and interviews with key educational staff members.

Annual reports dating back to 1991 from each of the three pathways were reviewed and summarized by task force members. Surveys for both students and faculty were developed, pilot tested, and summarized by task force members with help from OAS consultants. Surveys were administered in spring quarter of 1997 to all students. First- and second-year students were surveyed in their respective class locations, as a group; third- and fourth-year students received paper copies in their college mailboxes. Return rates were much lower for clinical-year students due to clinical assignments and time of the survey. Faculty surveys were distributed through internal mail services to faculty with 50–100% academic appointments. Likert-type survey items were analyzed using descriptive statistics: frequencies, percentages, cross-tabulations, means, and standard deviations. For reporting purposes "very satisfied" and "satisfied" were combined into "satisfied," and "very unsatisfied" and "unsatisfied"

were combined into "unsatisfied." Documents, interview notes, and other qualitative data were analyzed using domain analysis of key words and phrases.[20]

Results

Academic Outcomes. No difference across pathways was observed for graduation rates or grades on clinical rotations, but more IS students were in Alpha Omega Alpha (24% IS, 17% LD, 14% PBL) and higher percentages of both IS and PBL students received more departmental awards than did LD students.

Student Survey. The student survey was designed to learn how students choose their pathways and assess their satisfaction with their choices. The students were also asked to comment on their impressions of all three pathways.

Of the 839 student surveys distributed, 467 usable responses were returned (55.6%). The return rate was biased toward the basic science classes (year one = 92%, year two = 76%, year three = 43%, and year four = 11%). Return rates by basic science pathway for each class surveyed resembled the proportion of students enrolled across pathways (LD = 69%, PBL = 17%, and IS = 12%). Because so few fourth-year students returned the survey, their data were not used.

Having a choice of pathways was a significant factor in the students' decisions to come to the college: 56% of the respondents agreed that choice of basic science pathway influenced their decisions to attend the school.

Based on the students' responses, the factors that contributed to a student's choice of pathway were learning style, experience with nontraditional learning methods, personal and family needs, and needs for socialization. Sixty-two percent indicated that the LD pathway was their first choice. Many students stated a preference for it because it is a method with which they were familiar. Some felt that because of perceived weaknesses in their basic science backgrounds they needed the structure provided by LD. Social factors that contributed to pathway choice were distance from campus, need for contact with students and teachers, and need to make friends and network.

The PBL is the only pathway that caps enrollment at 35. Twenty-eight percent of the survey respondents (131 students) identified the PBL as their first choice of pathway; of these, nearly 40% (52 students) matriculated into other pathways. Students stating preferences for the PBL said that they either had had experience with group work in the past or believed that through PBL they could learn clinical reasoning skills early.

Nine percent of the respondents identified IS as their first choice. However, 12% reported participating in the IS pathway. Some students from the PBL wait list had chosen the IS pathway once it was determined that they would not be admitted into the PBL pathway. The students who chose IS as their first choice cited the flexibility of the pathway as their primary reason. This pathway tends to attract more nontraditional students such as older students with families, married students, or students interested in the MD–PhD program. Many stated that they would not have been able to complete medical school without the flexibility offered by the pathway. Others appreciated the opportunity to manage their own time by either accelerating or decelerating their pace through the basic sciences.

Overall, student satisfaction with their basic science pathways was high: almost 82% were satisfied with their pathways; only 9% reported being unsatisfied. Across the three pathways, PBL students reported being the most satisfied (91%), and 93% of the PBL students would have chosen it again. The IS students were almost as satisfied with their pathway, with 86% reporting satisfaction, although only 76% said that they would choose it again. The LD students were the least satisfied, with 79% stating that they were satisfied and only 63% said that they would choose that pathway again. No difference across cohorts was observed. The proportion of students expressing a preference for a given pathway was the same for each class: 42% said that they would pick LD, 41% said they would pick PBL, and 17% said they would pick IS.

Overall, 52% of the students felt they had missed something offered by the other pathways (54% of LD, 41% of PBL, 51% of IS students). Many LD students felt that they missed the clinical experience, case studies, and active learning that was offered by the PBL. On their own initiative, non-PBL students have started a case-study interest group in an effort to make up for this perceived need. Alternative-pathway students felt that they missed out on well-presented and organized material from content experts, comprehensive coverage, pressure to perform, and proper pronunciation of medical terms.

The overwhelming response by students was that choice was very important and that students have different learning styles. They felt that choice attracts a higher caliber of students and shows that the school is a progressive medical school. Over 90% of the respondents agreed that the school should continue to offer multiple basic science pathways.

Faculty Survey. All 568 faculty with 50% or greater appointments were surveyed; 133 (23.4%) responded. Of the 133 respondents, 23% were from basic science departments, 48% from clinical sciences, and 29% did not provide their departments.

Nineteen percent of the respondents reported no teaching experience in any pathway. Sixty percent taught in only one pathway (LD 50%, IS 1.5%, PBL 7.5%). Fourteen percent of the respondents reported experience in two pathways (LD/IS 4.5%, LD/PBL 7.5%, IS/PBL 2.3%). Seven percent participated in all three pathways.

The faculty respondents were generally satisfied with their student interactions in each pathway (54% of LD, 53% of IS, and 87% of PBL faculty). The basic science and clinical faculty disagreed on the appropriateness of the distribution of their teaching, research, and service time: 80% of the basic science faculty were satisfied with the time distribution, while only 47% of clinical faculty were satisfied.

When asked, "In your opinion is it important that the College of Medicine and Public Health continue to offer three preclinical pathways?" the faculty responses of those who expressed an opinion were split almost evenly (38% yes, 39% no, and 22% no opinion). For the faculty who identified their departments, approximately half replied in the affirmative (47% of basic science faculty, 50% of clinical faculty), and 19% had no opinion.

Discussion

Based on student and faculty opinions from surveys and comparison of pathway outcomes for 1993 to 1997, the task force unanimously recommended that the college maintain three basic science pathways. The presence of three preclinical pathways provides the college tremendous flexibility to accommodate student learning styles and time requirements. Students highly value the commitment of the college to medical education by accommodating their different student learning styles. Providing three pathways is also an important factor in the recruitment and admission of high-quality students. Differences in outcome measures are small and may be attributed to higher pre-matriculation statistics for IS and PBL students.

The three basic science pathways are important in maintaining the positive image of medical education at the college. This is true both for current medical students and for those applying. Requests for the PBL pathway from entering students averaged 46% of the entering classes of 1994–1997, and IS enrollments have increased dramatically. Faculty are generally satisfied with student interactions in the LD and IS pathways, but are most satisfied with their interactions with the PBL students.

130

Three pathways provide for differences in learning styles, as well as offering time for independent learning, research, and outside interests. Time flexibility by pathway is greatest with IS, followed by PBL, and least with LD. Student satisfaction with their current pathways is very high: 91% of PBL, 86% of IS, and 79% of LD students were satisfied with their basic science pathways. In spite of the high satisfaction levels, however, approximately half of the students felt that they had missed something in their pathways that was available in another pathway. Student comments indicated that this lack was not one of content material but rather in the social and pedagogic opportunities with faculty and other students. Eighty seven percent (87%) of the students agreed that the college should continue to offer three basic science pathways; only 5% disagreed.

Low faculty response rates, lack of teaching experience in the pathways, and "no opinion" responses make it difficult to interpret the faculty survey data. Therefore, the task force recommended educating faculty about the importance of the three pathways and their recruiting and retention benefits.

Conclusions

The Ohio State University College of Medicine and Public Health is well served by offering three parallel but alternative basic science curricula and will continue to do so. The large entering class size (210) and the three different pathways make the college fertile ground for comparison of alternative curricula. This study confirms previous conclusions in the literature about independent study and problem-based medical education in terms of outcomes, flexibility, choice, and student and faculty satisfaction and preferences. In addition, it established that multiple curricula are important factors in admissions, educational reputation, and accommodating various student learning styles.

References

1. Way DP, Biagi B, Clausen K, Hudson A. The effects of basic science pathway on USMLE Step 1 scores. Acad Med. 1999;74(10 suppl):S7–S9.
2. Herrick CA, Jenkins TB, Carlson JH. Using self-directed learning modules: a literature review. J Nurs Staff Dev. 1998;14:73–80.
3. Trzebiatowski GL, Williams JH, Sachs LA, Altman M, Bellchambers M. Independent study: 10-year programme review. Med Educ. 1987;21:458–63.
4. Huang AH, Carroll RG. Incorporating active learning into a traditional curriculum. Am J Physiol. 1997;273(6 Pt 3):S14–S23.
5. Todd KH, Braslow A, Brennan RT, et al. Randomized, controlled trial of video self-instruction versus traditional CPR training. Ann Emerg Med. 1998;31:364–9.
6. Zitzmann MB. Comparing the learning outcomes of lecture and self-instruction methods in a senior clinical laboratory science course. Clin Lab Sci. 1996;9:198–201.
7. VanArsdale SK, Hammons JO. Student-oriented learning outlines: a valuable supplement to traditional instruction. J Cont Educ Nurs. 1998;29:22–6.
8. Schlomer RS, Anderson MA, Shaw R. Teaching strategies and knowledge retention. J Nurs Staff Dev. 1997;13:249–53.
9. Randels PM, Kilpatrick DG, McCurdy L, Saunders PH. Comparison of the psychiatry learning system and traditional teaching of psychiatry. J Med Educ. 1976;51:751–7.
10. Gershen JA, Jedrychowski JR. The effect of supplemental lecture, evaluation method and instructional type on student performance in a pre-clinical technique project. J Dent Educ. 1979;43:276–80.
11. Graham HJ, Seabrook MA, Woodfield SJ. Structured packs for independent learning: a comparison of learning outcome and acceptability with conventional teaching. Med Educ. 1999;33:579–84.
12. Vernon DT, Blake RL. Does problem-based learning work? a meta-analysis of evaluative research. Acad Med. 1993;68:550–63.
13. Albanese MA, Mitchell S. Problem-based learning: a review of literature on its outcomes and implementation issues. Acad Med. 1993;68:52–81.
14. Saarinen-Rahiika H, Binkley JM. Problem-based learning in physical therapy: a review of the literature and overview of the McMaster University experience. Phys Ther. 1998;78:195–207; discussion 207–11.
15. Thomas RE. Problem-based learning: measurable outcomes. Med Educ. 1997;31:320–9.
16. Washington ET, Tysinger JW, Snell LM, Palmer LR. Implementing problem-based learning in a family medicine clerkship. Fam Med. 1998;30:720–6.
17. Finch PM. The effect of problem-based learning on the academic performance of students studying podiatric medicine in Ontario. Med Educ. 1999;33:411–7.
18. White MJ, Amos E, Kouzekanani K. Problem-based learning: an outcomes study. Nurs Educ. 1999;24(2):33–6.
19. Walton JN, Clark DC, Glick N. An outcomes assessment of a hybrid-PBL course in treatment planning. J Dent Educ. 1997;61:361–7.
20. Spradley JP. The Ethnographic Interview. New York: Holt Rinehart and Winston, 1979.
21. Patel VL, Groen GJ, Norman GR. Effects of conventional and problem-based medical curricula on problem solving. Acad Med. 1991;66:380–9.

131

The Health Sciences and Technology Academy: Utilizing Pre-college Enrichment Programming to Minimize Post-secondary Education Barriers for Underserved Youth

SHERRON BENSON MCKENDALL, PRISCAH SIMOYI, ANN L. CHESTER, and JAMES A. RYE

West Virginia is considered one of the most rural states in the nation, with over 60% of its population classified as rural.[1] The state experiences relatively high unemployment, and it ranks among the lowest (49th) of all states in median household income.[2] Fifty-eight percent of the students in West Virginia counties are eligible for free or reduced-price lunch.[3] Furthermore, only 14.7% of adult residents 25 years and over have attained a bachelor's degree or higher,[4] putting the state 50th in higher education.

The rural nature of the state coupled with economically depressed communities has limited the availability of secondary-level science courses required for health sciences majors in college. Additionally, most counties in West Virginia are considered medically underserved, and therefore it is important to increase the number of health care providers in rural areas of the state.[5] However, if the state's under-represented students do not receive adequate preparation in pre-college math and science, the proportion who can attend college and succeed will continue to be limited,[4] and the pool for the health professions will be too small.

To overcome some of these barriers, West Virginia University and 21 West Virginia counties have come together in the Health Sciences and Technology Academy (HSTA) in a community–campus partnership. Its web site is ⟨http://www.wv-hsta.org\⟩. A pre-college enrichment program,[6] HSTA helps students learn tools to enable them to progress through high school, college, and professional school. The HSTA program consists of an on-campus (WVU) Summer Institute at West Virginia University where students and science teachers are engaged in learning activities facilitated by science and education faculty. These science teachers also facilitate HSTA community-based science clubs during the school year. The HSTA model uses the inquiry-based theory that encompasses problem posing, problem solving, and persuasion.[7,8] Research suggests that inquiry activities emphasizing problem solving enhance middle-level students' self-confidence in mastering science and their attitudes towards the discipline.[9] Furthermore, inquiry-based learning is considered fundamental to students' understanding of science concepts and processes. The National Science Education Standards (NSES) call for greater emphasis on "inquiry into authentic questions generated from student experiences [which] is the central strategy for teaching science."[10] As a follow up to the NSES, a practical guide has been developed for educators who wish to emphasize inquiry-based instruction.[11] A principal thrust within the community science clubs is inquiry-based learning of science through extended investigations and community service projects.[12] The model also engages students in authentic learning processes (i.e., real-world problem-solving circumstances), which are both fun and challenging.[13,14] Students' projects often target health-related topics and may potentially inform and benefit various communities through dissemination at local and state levels.[15]

Methods

HSTA's effect on the academic success of its graduates and their decisions to pursue post-secondary studies and/or health sciences majors was assessed using quantitative and qualitative methods. HSTA participants are selected based on at least two of the following criteria: African American, financially disadvantaged, rural, and first generation aiming for higher education. Participants are admitted to HSTA during the ninth grade and participate in various activities until they graduate from high school, at which time they are considered HSTA graduates. There are 35 and 61 HSTA graduates for the 1998 and 1999 academic terms, respectively.

Telephone interviews were conducted in the fall and spring of the 1999–2000 academic term. Graduates were asked a series of questions that employed a Likert-type scale regarding HSTA's impact on pursuit of post-secondary study (1 = no impact to 5 = very high impact), choosing a health sciences major (1 = no impact to 5 = very high impact), preparation for college (1 = not at all prepared to 3 = extremely prepared), and preparation for major (1 = not at all prepared to 3 = extremely prepared). The participants were also asked to briefly explain why they had rated the program's impact and preparation levels as such.

In the fall of 1999, HSTA's impact on graduates' college performances was assessed with an independent t-test comparing the mean (μ) grade-point average (GPA) of 40 HSTA students (experimental) with that of 120 non-HSTA students at West Virginia University (WVU) in Morgantown, West Virginia. The 120 non-HSTA students were randomly selected from those enrolled at WVU with the same status (e.g. freshman), declared major, and residency. In order to achieve an effective sample size based on these characteristics (i.e., status, major, WV residency), three controls were matched to each experimental case.

Results and Discussion

Interviews. The return rates for questionnaires on pursuing post-secondary study and preparation level were 97% (93 students) and 80% (77) for the combined cohorts (1998 and 1999).

Post-secondary Study. The graduates' responses about HSTA's impact on their decisions to pursue post-secondary study indicate a strong impact, 3.88 and 3.96 (5 = very high) for the 1998 and 1999 graduates, respectively. The graduates provided a variety of reasons. One stated, "Before HSTA I didn't even know I could go to college because I'm from a poor family, and they gave me the chance to go to college." Another graduate affirmed that HSTA was the reason for her being in college. She posited that

> First of all, I didn't think I would be able to go to college and I really —I didn't have anybody in my family that said okay here is where I'm going, you ought to check this stuff out. When I came up here, I fell in love with the [WVU] campus. . . . . And I got in and it's nice to have contacts . . . . I knew a lot of the teachers and a lot of the faculty through HSTA and it really helped me out a lot.

A graduate who intended to pursue a nursing career reported that "when I was in high school and worked with HSTA for the summer, we got to work with the cadavers. That's hands-on experience that I'd never have had." Essentially, HSTA provides students with tangible experiences that bring excitement to the learning process. Not only is the program a pipeline for participants who wish to pursue post-secondary study, it also provides financial support for students who would not have had the opportunity to attend college.

College Major. Approximately 66% (23) of the 1998 graduates and 80% (49) of the 1999 graduates chose health sciences majors.

132

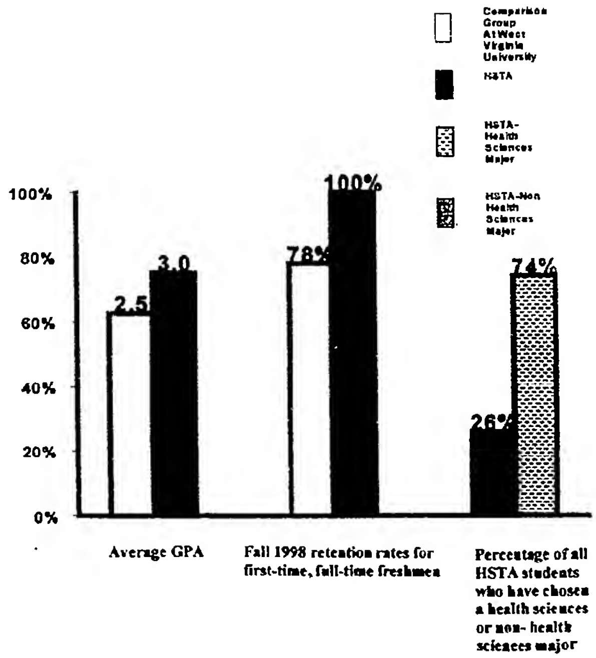

Figure 1. The Health Sciences and Technology Academy (HSTA) graduates' GPAs, their retention rates at West Virginia University compared with a non-HSTA group at the university, and their choices of major in higher education, fall and spring 1999–2000.

Among these graduates, the impact of HSTA on this decision ranged from moderate to high, 3.60 (1998 cohort) and 3.74 (1999 cohort) (5 = very high). The graduates rated the program highly because of the hands-on learning experiences it had afforded them. For example, one graduate stated

> Whenever we ... would do hands on experiences, it just made me more interested, especially in Psychology because when we go to mess with the brains ... it just made me more interested. It made it not seem as hard or as bad as what people think it is.

Another graduate reported that, "HSTA allowed me to see a lot of different areas in the health field that I wouldn't have seen otherwise. It kind of gave me a taste of everything and just sort of oriented me." Overall, these experiences not only expose students to various occupations of which they would otherwise have no knowledge, but it provides the opportunity to explore horizons within the realm of health sciences.

College Preparation. The graduates responded positively regarding level of preparation for college and major as a result of their participation in HSTA. Of both cohorts, 98% (94) were pursuing post-secondary study. In response to questions about college preparation, the mean responses were 2.45 and 2.46 (3 = extremely prepared) for the 1998 and 1999 graduates, respectively.

The graduates rated the program's preparation for their majors as 1.95 (1998 cohort) and 2.27 (1999 cohort) (3 = extremely prepared). Thus, the overall perception is that HSTA prepared them at least moderately for their majors. The higher rating given by the 1999 graduates may be due to the higher percentage of them who intended to pursue health sciences majors.

College Performance. The HSTA graduates had a significantly higher undergraduate GPA (population mean [μ] undergraduate GPA of 3.00) than the non-HSTA control group's mean GPA of 2.51. An independent *t*-test comparing the mean GPA of HSTA graduates with that of non-HSTA students at West Virginia University proved that there is a statistically significant difference between the GPAs ($\alpha = .05$, $p = .0014$, $t = 3.2495$). The result exemplifies HSTA's impact on those who matriculate to and graduate from the program. After the *t* test was performed, a 99% confidence interval of the true μ of the non-HSTA population was determined. The true μ of the non-HSTA population is between 2.31 and 2.71. Thus, we are 99% confident that the true mean GPA for the non-HSTA student population lies in the interval [2.31, 2.71]. Therefore, the true μ GPA (3.00) of HSTA students who attend WVU is not only higher than that of the control group (2.51) but also higher than that of the total non-HSTA population.

Retention. All HSTA's graduates who enrolled at WVU during the fall of 1998 were retained, compared with a rate of 78% for non-HSTA first-time freshmen (see Figure 1). Furthermore, an overwhelming 74% of 1998 and 1999 HSTA graduates are pursuing health sciences majors, compared with 26% of the graduates who have chosen other fields of study (see Figure 1). The graduates majoring in health sciences are particularly drawn to fields such as biology/chemistry, nursing, psychology, and allied health (see Figure 2).

Conclusion and Implications

The Health Sciences and Technology Academy provides a pipeline for underrepresented youth to pursue their higher education goals. Through pre-college enrichment measures, HSTA gives students multifaceted opportunities for academic enrichment, which help them to realize that they can become accomplished individuals in their communities and in society at large. Pre-college programs such as HSTA can provide enriching experiences for underprivileged students who may not foresee the importance of completing high school and going to college.

Although HSTA has not provided its graduates with enrichment

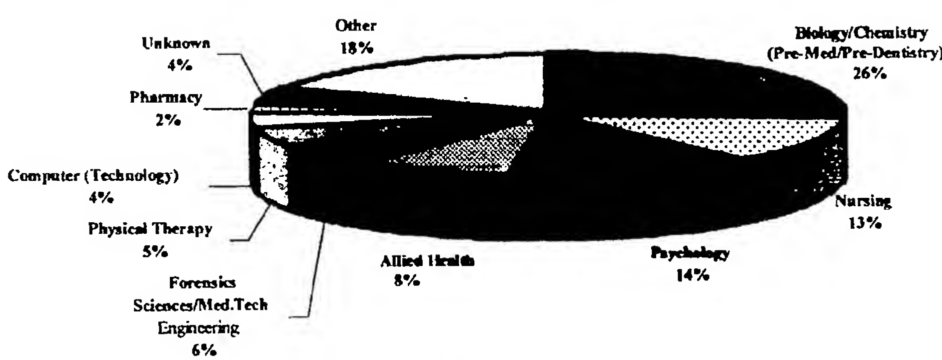

Figure 2. Percentages of Health Sciences and Technology Academy (West Virginia University) graduates choosing different higher education majors, 1998 and 1999.

133

experiences beyond high school, it can be assumed that the success of these students at WVU can be attributed, in part, to the pre-college enrichment provided by HSTA. Many of the graduates at WVU and other higher education institutions express a deep sense of fulfillment as a result of their participation in HSTA. Furthermore, many have expressed that their desires to pursue health sciences as well as technologic careers are due, in part, to the HSTA program. Their performance relative to that of non-HSTA students with similar interests is extremely encouraging. We believe that the HSTA model provides an exciting opportunity to extend inquiry-based learning, via longitudinal science projects, beyond what otherwise would be possible in the science classroom. All evidence indicates that the long-term benefits of this pre-college enrichment program will be positive.

Correspondence: Sherron Benson McKendall, PhD, Health Sciences and Technology Academy, PO Box 9026, Robert C. Byrd Health Sciences Center, Morgantown, WV 26506. Reprints are not available.

References

1. Table 1. Urban and Rural Population: 1990 to 1990. ⟨http://www.census.gov/population/censusdata/urpop0090.txt⟩. Accessed 3/99. Washington, DC: U.S. Census Bureau.
2. Table D. Median Income of Households by State: Income 1998. ⟨http://www.census. Gov/hhes/income/income98/in98med.html⟩. Accessed 3/99. Washington, DC: U.S. Census Bureau.
3. West Virginia Child Nutrition Program: Percentage of Needy Students (By School) October 1999. Office of Child Nutrition, West Virginia Department of Education, Morgantown, WV.
4. Table B–13. Educational Attainment of Persons 25 Years Old and Over, for States: March 1997. Washington, DC: U.S. Department of Education, National Center for Education Statistics. Digest of Education Statistics, 1998.
5. Rye JA, Chester AL. WVU–community partnership that provides science and math enrichment for underrepresented high school students. Acad Med. 1999;74:352–5.
6. Piehl DH, Mihm JC. Winning pre-college education programs. Res Technol Manage. 1995;38:34–40.
7. Hinman RL. Content and science inquiry: what inquiring minds *need* to know. The Science Teacher. 1998;65:25–7.
8. Peterson N, Jungck J. Problem-posing, problem-solving, and persuasion in biology. Acad Comput. 1988;2:14–14, 48–50.
9. Helgeson S. Research in problem solving: middle school. In: Gabel D (ed). Handbook of Research on Science Teaching and Learning. New York: MacMillan, 1994, 248–68.
10. National Research Council. National Science Education Standards. Washington, DC: National Academy of Sciences, 1996:31.
11. Olson S, Loucks-Horsley S. (eds). Inquiry and the National Science Education Standards: A Guide for Teaching and Learning. ⟨http://www.nap.edu/catalog/9596.html⟩. Accessed 5/00. Washington, DC: National Academy Press, 2000.
12. Rye JA, Bardwell G, Hu J. Connecting science, health, and technology through authentic investigations. Science Educator. 1999;8:19–24.
13. Cronin JF. Four misconceptions about authentic learning. Educational Leadership. 1993;50:78–81.
14. Gordon R. A curriculum for authentic learning. Education Digest. 1998;63:4–9.
15. Magnusson SJ. The learning environment as a site of science education reform. Theory Into Practice. 1995;34:43–50.

-134

The Mount Sinai Humanities and Medicine Program: An Alternative Pathway to Medical School

MARY R. RIFKIN, KENNETH D. SMITH, BARRY D. STIMMEL, ALEX STAGNARO-GREEN, and NATHAN G. KASE

In 1984 the AAMC report of the Panel on the General Professional Education of the Physician[1] recommended that students preparing for medical school should strive for a curriculum that provides a broad study in both the sciences and the humanities and that required courses should be kept to a minimum. One way to encourage premedical students to follow a truly broad liberal arts education would be to accept students to medical school early in their college careers, thereby alleviating the pressure to focus excessively on the traditional science-based curriculum. Because there is no evidence to suggest that science majors are necessarily more qualified for medical school, we initiated an experimental program that encouraged humanities and social science majors to pursue their individual interests in college and to obtain a broad, maturing, liberal arts education. Such students might be expected to be less focused on the technology of medicine, bring different perspectives to the practice of medicine, and simultaneously diversify the student body.

In 1989 the Mount Sinai School of Medicine (MSSM) started the Humanities and Medicine (H&M) Program, an early-assurance-of-admission program designed for humanities and social science majors at a targeted group of five liberal arts colleges and universities (Amherst, Brandeis, Princeton, Wesleyan, and Williams).[2] Students in this program are selected during the first semester of their sophomore year in college. Admission into the program is based on a written application with personal essays, verbal and math SAT scores, high school and college transcripts, letters of recommendation, and personal interviews. The students are required to major in the humanities or social sciences and are required to complete only one year of college biology and one year of college chemistry with a grade of B or better.

Admission to MSSM is contingent upon successful completion of undergraduate studies, provided the GPA does not drop below a minimum of 3.0. MCAT scores are not required. In addition, students are required to spend an eight-week summer term at Mount Sinai after their junior year, during which they are exposed to clinical activities and complete a much abbreviated course on the principles of organic chemistry and physics relevant to medicine. Housing and a stipend are provided. Students admitted to the H&M program are under no obligation to attend Mount Sinai should their career choices change or another medical school appear more attractive. Also, the students have the option of deferring their admission to medical school for one year after obtaining the undergraduate degree.

This study reports the outcomes of ten years' experience with the H&M Program. Our experience shows that although students in this program have more academic difficulties in the preclinical years, they excel in the clinical/community setting and have greatly enriched the medical school environment. This program demonstrates that success in medical school does not depend on a traditional premed science curriculum.

Method

The achievements of all H&M students ($n = 85$) matriculating at MSSM between 1991 and 1997 have been compared with those of two matched cohorts of students who had been accepted through the standard admission process and had completed all standard premed science requirements. Students in each cohort were matched to the H&M students on the basis of year of matriculation, gender, age (within three years), category of educational institution (top 30 liberal arts colleges or universities, taken from the 1998 *US News & World Report* Survey), and, when possible, ethnicity, and were either humanities/social science majors or science majors. The groups of 85 students included students at different stages of their medical school careers and five classes of graduates (1995–1999).

For each group, academic performance in medical school in both basic science courses and clinical clerkships and performance on the USMLE Step 1 examination were analyzed. In addition to these quantitative indicators of performance, we performed an analysis of the students' overall medical school achievements and contributions to the medical school environment in terms of extracurricular activities, student leadership, and service, by evaluating their election to AOA and receipt of special awards. *P* values were determined using the χ^2 test.

Results

The undergraduate science/math background of students entering MSSM through the H&M program consists of one year each of biology and chemistry and a short summer course at MSSM, "Physics and Organic Chemistry Relevant to Medicine." This differs from the premed science/math requirements for all other students matriculating at MSSM, namely one year each of biology, chemistry, organic chemistry, physics, and math. The data in Table 1 show that a significantly higher proportion of H&M students had at least one course failure in the basic science years than did the students with traditional premed science backgrounds, who were either humanities majors or science majors. Over 75% of the course failures of H&M students occurred in the first semester of year one, where there were nine failures in biochemistry, six in embryology, six in cell biology, and five in gross anatomy (data not shown). Among the 20 H&M students who failed one or more courses, nine students failed multiple courses, with the range being up to four courses. In the second basic science year, the proportion of H&M students with at least one course failure decreased, with no single course having a disproportionate number of failures.

Compared with their classmates, the H&M students had a higher failure rate on the USMLE Step 1 examination (Table 1), although all these students eventually passed it (data not shown). In an attempt to determine whether failure on the Step 1 examination could be predicted from data available at the time of acceptance into the H&M program, we analyzed the correlation of these students' SAT scores with their performances on the Step 1 examination. Neither Verbal SAT ($R^2 = 0.08$) nor Math SAT ($R^2 = 0.07$) scores correlated with the Step 1 examination score. However, all students who failed the Step 1 examination had Verbal SAT scores ≤ 650.

In the clinical years of medical school, statistically significant differences in performance between the H&M students, when compared with the matched cohorts, were less evident. The failure rate of H&M students in clinical clerkships (Table 1), the garnering of clerkship honors, and election to AOA (Table 2) were not significantly different from those of either matched humanities majors or matched science majors. In fact, the H&M students with mul-

135

TABLE 1. Performance of Humanities and Medicine Students Compared with Two Matched Cohorts in Preclinical Courses, in Clinical Clerkships, and on the USMLE Step 1 Examination, Mount Sinai School of Medicine, 1991–1998

	Humanities and Medicine Students	Matched Regular Premed Students, Humanities Majors	Matched Regular Premed Students, Science Majors	*p*
Basic science year one: students with at least one course failure	20 (85)*	11 (85)	2 (85)	<.001
Basic science year two: students with at least one course failure	10 (76)	3 (77)	3 (77)	<.03
Clinical clerkships: students with at least one clerkship failure	6 (76)	2 (77)	1 (77)	< 09
USMLE Step 1: students failing on first try	10 (76)	2 (77)	3 (77)	<.02

*Number in parentheses indicates total number of students analyzed.
p values were determined by chi-squared analysis of the data.

tiple clerkship honors were often the same students who had had academic difficulty in the basic science years or who had failed the Step 1 examination. Analysis of specific clerkships indicated that the H&M students excelled in the psychiatry and pediatrics clerkships (data not shown).

In the preclinical years, Book Awards are given to those students who have performed outstanding extracurricular activity within the community or who have contributed time and energy in service to the institution. Over half the Book Awards were awarded to H&M students (Table 2). H&M students are also disproportionately represented on various subcommittees of the Student Council and other institutional committees, as well as serving in large numbers as student group representatives to national organizations such as the American Medical Student Association, American Medical Women's Association (AMWA), and Students for Equal Opportunity in Medicine (SEOM). Furthermore, a greater proportion of H&M students than students in the two matched cohorts received prizes and awards at graduation (Table 2).

Additional data, not shown, indicate that the H&M students completed medical school at the same rate and did not have a higher attrition rate than students entering medical school with more traditional premed backgrounds. Analysis of residency placements indicated that 77% of the H&M students placed in university hospital-based programs, as opposed to affiliate hospital-based programs, as did 74% of the science majors cohort and 69% of the humanities majors.

Discussion

The Humanities and Medicine (H&M) Program challenges the long-standing belief that there is a necessary relationship between undergraduate science preparation and the successful completion of medical school and physician excellence. Students in this program are encouraged to use their time in college to pursue in depth their individual interests in their particular majors, which must be in the humanities or social sciences. They often spend considerable time in study abroad, independent research projects in their major fields, or extracurricular activities on campus, such as creative or performing arts or journalism. These students thereby avoid premature specialization and can obtain a broad, maturing, liberal arts education.

The academic performance of H&M students at MSSM has been compared with the performances of two matched cohorts: matriculated students with the standard, required science course background who majored either in the humanities/social sciences or in science. Since the medical school basic science courses are all graded by a norm-referenced rather than criterion-referenced system, and all the other students had had at least two more years of science, including organic chemistry, it is not surprising that the H&M students had more academic difficulties in the preclinical years than did the traditional premed students. However, in the clinical years and in the community setting, the H&M students were similar to the traditional premed students in garnering clerkship honors, institutional awards and prizes, and election to AOA.

TABLE 2. Numbers of Honors and Awards Given to Humanities and Medicine Students and to Two Matched Cohorts, Mount Sinai School of Medicine, 1991–1999

	Humanities and Medicine Students	Matched Regular Premed Students, Humanities Majors	Matched Regular Premed Students, Science Majors	*p*†
Clerkship honors/students				
0 honors grade	7	11	16	.12
1–5 honors grades	57	51	44	.06
6–10 honors grades	12	15	15	.77
Alpha Omega Alpha	14 (76)*	9 (77)	15 (77)	.08
Book awards				
First year	5 (85)	2 (85)	1 (85)	.21
Second year	9 (77)	2 (77)	4 (77)	.06
Graduation awards/prizes, classes of 1995–1999	21 (61)	10 (61)	13 (61)	.03

*Number in parentheses is total number of students.
†*p* values were determined by chi-squared analysis of the data.

All the H&M students who failed clinical clerkships ($n = 6$) also had course failures in *both* of the basic science years, whereas none of the students in the cohort groups ($n = 3$) who failed clinical clerkships had course failures in both of the first two years of medical school. While the numbers are small, these data, together with other information about these students' career goals and motivation, suggest that this subset of H&M students may represent students not wholly committed to the study of medicine. There was no evidence in the undergraduate records of these students that could have predicted this pattern of failure.

Although previous reports by others[3,4] indicate that there is no significant correlation between medical school performance and undergraduate major, the students in those studies had completed the required science courses of a traditional premedical undergraduate education. Our report on the performance of the H&M students, who have majored in the humanities or social sciences and who have had minimal science education in college, indicates that, as might be expected, these students have significantly more academic difficulty in the basic science years in medical school than matched classmates who have completed the traditional premedical curriculum. Moreover, we found that all H&M students who failed the USMLE Step 1 exam had verbal SAT scores equal to or less than 650. Thus, in an effort to minimize the number of students whom we might predict would have difficulty in medical school, we have decided to pay particular attention to the verbal SAT score in our admission process, as well as to scrutinize applicants' high school science and mathematics achievements with care.

The premise on which the H&M Program is based is that by eliminating the requirement for traditional premed requirements in college, students have more time to devote to their humanities majors and other pursuits and thus have time to broaden their backgrounds, which would be beneficial to their careers as physicians. These students bring to the medical school certain qualities and outlooks that positively impact the entire medical school community. They have been among the founders of various musical ensembles, theater groups, and art exhibitions, as well as members of the executive board positions of MSSM chapters of AMWA and SEOM. The first woman president of the Student Council was an H&M student. There is no doubt that the MSSM community has been enriched by the diversity of interests brought to the campus by the H&M students.

The studies reported here should lead us to reconsider the need for the traditional science courses as a prerequisite for success in medical school. Numerous published reports[5-7] have questioned the emphasis on science knowledge in the selection of medical students and have suggested that studies in the humanities may enhance effective patient interaction and communication. By selecting highly qualified, intelligent students early in their college careers and allowing them to develop their curiosity in their chosen fields of interest, as well as involving themselves in community and extracurricular affairs, we have shown that such students successfully complete medical school and excel in clinical activities. We intend to track these students as they complete their residencies and establish their careers to be able to more fully evaluate their contributions.

Correspondence and requests for reprints: Mary R. Rifkin, PhD, Mount Sinai School of Medicine, Box 1475, New York, NY 10029.

References

1. Muller R. (chair). Physicians for the Twenty-first Century: Report of the Panel on the General Professional Education of the Physician and College Preparation for Medicine. Washington, DC: Association of American Medical Colleges, 1984.
2. Stimmel B, Smith K, Kase N. The Humanities and Medicine Program: the need for the traditional premedical requirements. Acad Med. 1995;70:438.
3. Ashikawa H, Hojat M, Zeleznik C, Gonnella JS. Reexamination of relationships between students' undergraduate majors, medical school performances, and career plans at Jefferson Medical College. Acad Med. 1991;66:458–64.
4. Zeleznik C, Hojat M, Veloski J. Baccalaureate preparation for medical school: does type of degree make a difference. J Med Educ. 1983;58:26–33.
5. Dickman RL, Sarnacki RE, Shimpfhauser FT, Katz LA. Medical students from natural science and nonscience undergraduate backgrounds. JAMA. 1980;243:2506–9.
6. Doblin B, Korenman S. The role of natural science in the premedical curriculum. Acad Med. 1992;68:539–41.
7. Neame RLB, Powis DA, Bristow T. Should medical students be selected only from recent school-leavers who have studied science? Med Educ. 1992;26:433–40.

137

1999 JACK MAATSCH MEMORIAL PRESENTATION

The Epistemology of Clinical Reasoning: Perspectives from Philosophy, Psychology, and Neuroscience

GEOFFREY R. NORMAN

Physicians' clinical reasoning has been an active area of research for about 30 years. The goal of the inquiry has been to reveal the processes whereby doctors arrive at diagnoses and management plans (although as Elstein correctly points out in his discussion of this paper,[1] the focus has been more on the former than on the latter) so that we could use this information to devise specific instructional strategies or support systems to make the acquisition and application of these skills more efficient and effective. Initially, these "clinical reasoning skills" were conceived as general, and content-independent, so that they could be observed in all clinicians working through any problems. That is, they were thought of as a general mental faculty, presumably rooted in the architecture of the mind, which would be brought to bear on solving clinical problems.

However, the research findings did not support this viewpoint. Elstein and Shulman[2] showed that whatever clinical reasoning was, it was definitely not skill-like, in that there was consistently poor generalization from one problem to another, a finding that ultimately sounded the death knell for evaluation methods such as patient management problems. The past 30 years have seen an accumulation of evidence, in medicine and many other disciplines,[3] about the nature of the process, and shown the importance and centrality of knowledge. The central issue of this revised research program is achieving an understanding of how knowledge is initially learned, how it is organized in memory, and how it is accessed later to solve problems.

A second research program in medical decision making also emerged from research of the early 1970s. As Elstein discusses in the companion paper, this program "views diagnosis making as opinion revision with imperfect information."[1] From the decision-analytic perspective, the best decisions arise from the application of a statistical decision rule to data; any other method is suboptimal. Thus, the research agenda is directed to identifying areas such as medicine where humans function in a suboptimal way, and attempting to understand the strategies, the heuristics and biases, they apply to arrive at these suboptimal decisions.

Elstein states that "it seems to me that decision theory is at least as promising as the study of categorization processes." He may well be correct. But the two schools highlight a fundamental epistemologic dilemma that the remainder of this paper addresses: Will we understand more about the nature of clinical diagnosis by focusing on the diagnostician and striving to understand the mental processes underlying diagnosis, or by focusing on the clinical environment and attempting to understand the statistical associations among features and diseases? To what extent is the world of clinical reasoning "out there" and comprehensible by understanding the relation between symptoms and diseases, and to what extent is it "inside" and understandable only by examining mental processes in detail?

Further dilemmas face us as we examine the research in clinical reasoning. "Organization of knowledge" is viewed as a critical determinant of expertise in medicine. But it is not really clear what is meant by organization of knowledge. Is knowledge organized hierarchically with general concepts at the top, more specific scripts in the middle, and specific instances at the bottom?[4] Is it organized in networks with nodes and connections,[5] as a symptom-by-disease matrix,[6] as propositions with causal links,[7] as collections of semantic axes,[8] or as individual examples with no overarching concepts, as some of my earlier research claimed?[9]

A perusal of these various studies leaves the reader with only one overall impression—that the human mind is incredibly flexible and can organize and reorganize information at will and seemingly effortlessly to give the researcher exactly what he or she wants to hear. It is no coincidence that propositional networks are disturbingly idiosyncratic and not apparently reproducible.[5] My view is that all of these concept architectures are produced on the fly at retrieval, in order to satisfy the expectations of the researcher, and none can claim special status as *the* way knowledge is organized. Do you want the clinician to tell you the probability that myocardial infarction (MI) will present with referred pain to the back? Can do. The nature of the neural pathways linking the heart and the upper arm? Sure. The hair color of the last patienr they saw with an MI? Red. Given this incredible diversity of knowledge from specific to general, it seems likely that any attempt to uncover a representation of knowledge consistent with a particular perspective from fairly directive probes will be successful; however, the ultimate form of this knowledge (if that is even an issue worth addressing) will remain elusive.

Still, if the clinician's mind is really that malleable, then this poses a serious challenge to the research tradition. Are there really any more "basic" or "primitive" forms of knowledge? How can we understand the nature of clinical reasoning if it appears to be this flexible? These were the questions that presented themselves as I reviewed the studies of clinical reasoning. As I thought about these issues, I began to explore other perspectives on the nature of knowledge and knowing from philosophy, psychology, and neuroscience, and started to identify common threads that, I think, can shed some light on these questions. As I did so, I found myself moving back and forth among three kinds of knowing, more or less from specific to general:

1. How does the clinician come to know about diseases? How might diseases be represented in his or her mind?
2. How do we as researchers come to understand domains of science, whether these are the diseases of clinical research or the workings of the clinician's mind?
3. What do we mean by knowing? What do we mean when we say we understand something?

In the remainder of this article I roam freely among these levels, since many of the writings I uncovered inform all levels. But I must begin with a disclaimer. My journeys in this field are as an amateur, and are recent. I have been heavily influenced in my interpretations by two books. The first is *Lessons from an Optical Illusion*, by Hundert,[10] who took the brave step of trying to find links among philosophy, psychology, and neuroscience. His goal was to place ethics in a context of these disciplines; mine is to turn these general truths to an understanding of clinical reasoning. A second major influence on my thinking is a book called *What is this Thing Called Science?*

by Chalmers[11]—a wonderful and readable review of classical philosophy and philosophy of science. I highly recommend both.

The starting point of my discourse is a critical examination of the concept of disease. My intention is to use the exploration of disease as a case study of how we come to know about things.

What Is a Disease?

Through advances in biology, physiology, and molecular biology, we have come to a deep understanding of the mechanisms of many diseases. It seems almost nonsensical to now turn the clock back and ask what a disease is. But this small departure may serve us in good stead in understanding better what a concept is and how people identify concepts.

Let's take two examples:

- **Is syphilis a disease?** Absolutely. It fits the medical model to perfection. A bacterium invades the host, stimulating a diversity of processes that ultimately are manifested in clinical signs. Osler said "understand syphilis and you understand all of medicine." But there is a small historical glitch. Syphilis has been with mankind for millennia and the signs and symptoms were well established long before the bacterium was isolated.
- **Is heart disease a disease?** Yes. Put a label such as anterior myocardial infarction on it, and it looks even more like a disease. But likely we are all harboring the precursors of ischemic disease as cholesterol plaques slowly accrue in our arteries. So in a manner of speaking, the prevalence of heart disease approaches 100%. Can we then still speak of it as a disease? And by the way, although there are many risk factors for heart disease, there is no clear cause. The same is true for cancer. We can easily identify cancerous cells on pathology slides, and we can correlate the clinical course with the accumulation of malignant lesions, but we all have microscopic tumors in our thyroids, and a third of men who die of unrelated causes are found to have prostate cancer.

All of these things seem disease-like because we can "explain" them at some lower level—plaques, bacteria, malignant cells. But there are many other diseases listed in textbooks that have no clear causes, no microscopic correlates, no known mechanisms. And it is well to bear in mind that although anthropologists and historians have identified evidence of (for example) tuberculosis dating back several thousands of years, and although old writings in medicine clearly describe the symptoms and clinical course of tuberculosis, the cause, the tubercle bacillus, was identified, by Koch, only as recently as 1884, and effective therapy has been available only since the 1940s. So the existence of a causal mechanism is hardly sufficient to claim that something is a disease. More generally, it is likely that exceptions to any definition of disease will be common.

Campbell et al., in a classic article, "The Concept of Disease," reported presenting clinicians and lay people with a series of medical conditions and asking them whether or not they were diseases.[12] Perhaps not surprisingly, doctors were more prone than lay people to call things such as lead poisoning and tennis elbow diseases. But there was otherwise quite good concordance. Infectious diseases—malaria, tuberculosis, syphilis, polio—topped the list. Other common or serious medical problems—lung cancer, diabetes, multiple sclerosis, cirrhosis—came next. At the bottom were things such as hangover, senility, heatstroke, tennis elbow, and drowning, which had English, not Latin, labels. These authors concluded that the features that best predicted the labeling of a condition as a disease were that the condition (1) was associated with an abnormality of structure or function (i.e., it had a "cause") and (2) was likely to be treated by a doctor. The latter was the stronger determinant, but regrettably, this seems tautological. Since doctors are in the business of dealing with disease, describing a disease as something that doctors deal with does not, in my view, advance our understanding much.

Let us consider the first predictor for a moment. Arguably one simplistic but functional view is that if a condition simply represents a cluster of signs and symptoms (for example, carpal tunnel syndrome, low back pain) it is less disease-like. Presumably this reflects a concern that a condition's features and associations among the features may be an illusory correlation (which humans are particularly good at making)[13] and not "real." There is good reason for such a degree of skepticism. Historically, many syndromes that existed 100 years ago, such as self-pollution, have now disappeared, and there is every indication that many contemporary syndromes, such as chronic fatigue, sick-building syndrome, Gulf War syndrome, and the myriad health problems believed to be caused by breast implants may go the same way. Conversely, the ability to explain disease through some underlying mechanism lends authenticity to it. Angina becomes much more believable if we can find narrowing of the lumen of the coronary artery on angiography, even though the association with the clinical manifestations is weak.

The Role of Basic Science

If we view the identification of the features of a disease as analogous to the findings of an experiment (in this case, an experiment conducted by a malicious deity) then one basis for distinguishing a disease from a non-disease is the extent to which the features can be explained by a scientific theory. Thus the infectious diseases are explained by a noncontroversial, and historically verified, theory of host and parasite. Chronic diseases such as atherosclerosis are a bit less disease-like since the theory underlying them is less secure. And as we move to syndromes such as chronic fatigue syndrome, we are less inclined to view them as diseases because no satisfactory scientific mechanism has yet been found to explain their features.

Turning to clinical reasoning, investigators such as Schmidt[14] and Patel,[15] in studying the role of basic science in clinical reasoning, have found repeatedly that clinicians rarely invoke mechanistic explanations. But as Schmidt has shown, the fact that they need not invoke mechanisms does not mean that they do not know them—the knowledge is available but is only rarely used. As he describes it, the knowledge is "encapsulated." While basic science may play only a minimal role in day-to-day-practice, it is arguably the only, or at least the major, route to undersranding in this domain. Of course, basic science need not be restricted to biology. In the same way, the basic science of epidemiology was fundamental to understanding the transmission of AIDS, just as Snow in the 1880s understood the mechanism of cholera transmission (the London water supply) long before the bacillus was isolated.

I believe we can now posit an explanation for the paradoxical findings of Schmidt and Patel. In the normal course of events, clinicians making diagnoses deal at the syndrome level, where the nature of the causal mechanism is irrelevant. The history and physical exam are directed at revealing the syndrome-like manifestations, which then point to tests directed at the underlying processes, and therapy. The textbooks of clinical diagnosis for "old" diseases probably have not changed much since Osler's time. The signs and symptoms are pretty well what they have always been, although of course some historic scourges—smallpox, diphrheria, cholera—are now nearly unheard of in the West, and others, such as AIDS, have taken their place. But despite the changes in our understanding of disease, the clinician attempting to make a diagnoses is dealing almost exclusively at the syndrome level. Occasionally, some understanding of underlying processes may help to sort out some conundrum, but one suspects that clinicians appear rarely to use basic science simply because their investigations of history and physical are directed to labeling the syndrome. Clinical reasoning reverts to a historically earlier form of the disease, following the biologic dictum that ontogeny follows phylogeny—the fetus passes through all stages of evolution before birth.

Campbell[11] elaborated the notion of disease in philosophical terms, describing two basic positions: the "nominalist" perspective and the "essentialist" perspective. In the nominalist view, a disease is simply a collection of abnormalities that appear to arise together. Thus the historical diseases of dropsy, consumption, and plague were recognized long before any causal agent was detected, although etiologies (such as "bad humors") were advanced. Conversely, the essentialist perspective presumes that the signs and symptoms arise from pathologic processes that can be identified and hopefully rectified. While it is tempting to place these two views in a historical order, the contemporary examples we have discussed indicate that the two perspectives represent extremes on a continuum, which, as we shall see, has parallels in both philosophy and psychology.

What is a Concept? Lessons from Philosophy

We can make some general observations about the concept of disease. First, a disease, like any concept, does not exist entirely "out there" but rather, to some degree, is a mental construct. Second, the category or concept called "disease" is not an all-or-none proposition; rather, particular exemplars have different degrees of disease-ness. Finally, it is awfully difficult to devise an explicit rule to aid in distinguishing between diseases and non-diseases. A rule such as "diseases are what doctors deal with" works quite well but is singularly uninformative. And we sense, without proof, that any rule we may devise is not going to be coldly analytic, but must have sub-rules such as "the more Latinesque it is, the more disease-like it is." So ironically, while it is relatively easy to devise rules to determine whether someone has a particular disease (although I will go on to show that the rules are not the whole story), it is a lot harder to devise rules for the overarching category called "disease."

These issues are not at all specific to disease, but rather are part of a large body of knowledge extending in space across at least three disciplines—philosophy, psychology, and neuroscience—and in time as far back as Plato. To explore this further, I now venture (with considerable trepidation) into a more general inquiry into the nature of concepts. I begin by revisiting some philosophical views on the nature of concepts.

The origin of concepts has been, in some sense, a nature–nurture debate.[4] However, this argument has focused not on whether human traits are inherited or learned (the usual spin on nature versus nurture), but rather on whether categories or concepts such as beauty, disease, table, or tree exist "out there" to be learned by individuals as they develop and mature (which would suggest that an individual's knowledge is formed from experience [nurture]) or are essentially a product of the mind (we impose order and category boundaries where none exists, as a result of the biological structure of the mind [nature]). A casual reading of any philosophy textbook reveals that this issue has been a central concern through the ages of the great minds—Plato, Artistotle, Descartes, Hume, Kant, etc. Let us briefly review the historical debate in mainstream philosophy, with a view to showing how thinking in philosophy can help to frame our perspective on clinical reasoning.

Modern philosophy began with Descartes, who emerges as the ultimate skeptic, and whose views have retained central status as the universal straw man for all his successors. His famous statement "cogito, ergo sum" (I think, therefore I am) has been a lodestone for philosophers and t-shirt makers for three centuries. Regrettably, this idea has been almost universally misunderstood. Most interpret it as a statement of the ultimate rational man; our humanity is defined in terms of our capacity for rational thought. Unfortunately, the statement had a much more humble meaning for Descartes. In continuing to question whether one could justify any external reality, to devise any conclusive argument for the existence of objects such as dogs and tables, Descartes was led to the desperate conclusion that the only thing he could be really sure of was his own thoughts. I think, therefore I am.

The antithesis of this position was championed by the English empiricists Locke and Hume. Their view was that the mind was a *tabula rasa*, a clean slate on which one's experience with the world was written. This interpretation seems perfectly acceptable for sensory experience, but is more difficult to sustain for higher concepts such as causation, temporality, or, for that matter, disease. Hume's resolution was to suggest that these notions emerged as a result of experience.

Kant reframed the issue in a way that is central to our subsequent journey through psychology and neuroscience. He recognized that thoughts can occur only as products of interactions between the mind and the external reality of experience; we construct experience. He maintained a rigid boundary between those properties that our minds bring to experience (which are hardwired) and those that emerge from experience. He eventually created a list of 12 "primitives"—object, causation, temporality, and nine others—that he claimed the mind imposed on the world of experience.

Hegel went one step further and recognized that the external world can influence the categories and labels we apply. The categories themselves do not emerge from our minds, but are influenced by the objects of our perceptions. The mind is not simply a clean slate upon which all experience is written in coherent form (Hume); nor is it the case that there is no uniform order in the outside world and that all concepts are mental inventions (Descartes); nor finally does the mind impose fixed structure or constructs on sensory experience (Kant). Instead, the concepts and the content both grow and evolve ("become") as a consequence of the interaction between the individual and the environment.

Finally, in this century, Wittgenstein extended these ideas further. He proposed that not only are concepts not fixed, they also are not definable by any set of logical rules. In pondering even commonplace concepts such as "dog," he realized that any attempt to devise rules is doomed. A dog has four legs—but if one is amputated it's still a dog. A dog barks—except an Egyptian Basenji. A concept—whether an abstract concept such as truth or a mundane concept such as dog, fork, or tree—emerges as a matter of "family resemblance." Robins are more bird-like than penguins; malaria is more disease-like than alcoholism. Wittgenstein proposed that concepts or categories are derived from family resemblances, not from fixed sets of defining attributes.

Thus the philosophy of concepts evolved from a Cartesian view, which is entirely intra-psychic and questions any external reality, and an empiricist perspective that presumes that all order and concepts exist as natural categories to be discovered by the human observer, to a Kantian interaction, in which the mind provides the categories or concepts and the external reality provides the objects to fill the categories, to a Hegelian perspective, which is much more organic, and in which thoughts and concepts themselves evolve and change as a result of interactions with external reality. Ultimately, we reach the perspective of Wittgenstein, which places even fewer constraints on concepts, which are a matter of family resemblance and thus can be elaborated only through extensive experience with the world's families.

Applying these notions to clinical reasoning, philosophy presents a larger framework in which to view our dilemma in defining a disease. To the extent that a disease is a concept, philosophy buttresses the middle ground between the notion that diseases exist entirely "out there" only to be discovered and learned and the notion that they are probably simply mental constructs. We can then think of the concept of disease as arising from an interaction between the thoughts of the perceiver and regular aspects and associations of the environment. Further, some diseases, such as syphilis, are more central members of the family; others, including the syndromes, are more peripheral.

As we shall see, this formulation finds remarkable support in research in both psychology and neuroscience, to which I now turn.

What is a Concept? Lessons from Psychology

One division in psychology has been preoccupied with the same issue as the philosophers: how do people learn concepts such as table, dog, or truth? But instead of relying entirely on reason for understanding, psychology seeks evidence to understand how people create and learn concepts. Perhaps in the course of doing so, psychologists deliberately skirt some of the tough epistemologic issues that preoccupy philosophers. On the other hand, in my own reading, I was struck by how the one informs the other. A simple example:

The Müller Lyer illusion,[16] shown in Figure 1, is pretty well known to all. We see the one vertical element as being longer than the other. Even though we can measure them and show them to be the same, the illusion is inescapable—a fine example of how we impose order (sometimes biased order) on the external world. But psychologists have gone further with this illusion, and questioned precisely *why* it is an illusion. In the course of doing so, they provide a nice illustration of Hegel's interactive model of mind. One hypothesis is that it is an illusion because our minds are seeing it in three dimensions, so that the symbol on the left is seen as the outside corner of a wall nearest the viewer, and the one on the right is seen as the inside corner of a wall farthest away from the viewer. Although the two vertical lines are objectively the same size, since the one on the left is seen to be nearer than the one on the right, the right one is "actually" longer. Deregowsky[17] tested the illusion in Zulus, who spend their lives in round houses, and found that they did not see it as an illusion. So, it is not an illusion because our brains are "hardwired" to see it as such (unless Zulus have different hardwiring); it is an illusion because of the particular experiences we have had with the world. On the other hand, the illusion reminds us that our perceptions do not necessarily mirror reality, as they are also shaped by internal assumptions (in this case, about perspective and the inference of a third dimension from the two-dimensional representations on the retina) that sometimes lead us astray.

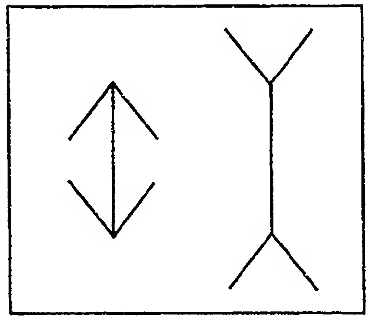

Figure 1. Müller Lyer optical illusion.

A second example from psychology leads us closer to our central concern with clinical reasoning. Most of us have, at one time or another, wondered whether the "red" we see is the same as the red seen by the person beside us. While the differences in perception are rarely likely to be as extreme as in the case of a childhood friend of mine whose color blindness was detected when he went to school and repeatedly drew green reindeer at Christmas, we have no real way of ever verifying the universality of "red." Is it just a linguistic device, or a cultural norm? After all, at some time we all had to learn, from our parents or friends, what red was. Perhaps it differs in different cultures. These questions, as they begin to cross the boundary between philosophy, psychology, and learning, are of more than passing interest.

Much of the fundamental work in concept formation has been done by Eleanor Rosch.[18] One area she studied was how colors are identified in different cultures. While, on the one hand, there appear to be small cultural differences in the boundaries between colors (e.g., the Navaho have only one word for blue and green (no wonder, with all that turquoise jewelry around),[19] Rosch showed that all cultures were unanimous in their choices of the best examples of red, yellow, or green. Even more interesting, Rosch discovered a primitive tribe, the Dani, who had words for "bright" and black only. She then taught them words for colors, using Dani words (e.g., tree) that were unrelated to color. One group learned the "primary" colors such as fire-engine red; the other learned Dani words for intermediate colors such as turquoise. The group learning red, yellow, and blue learned the associative words rapidly and effectively; the other group never did master the associations. Studies of this type provide support for the contemporary notion in philosophy that categories and concepts derive from our experience of the world; indeed there is surprising uniformity to these concepts in precisely those areas where we might expect that experience (such as the experience of color) is also universal.

Prototype theory was perhaps the first theory of concepts to be seriously applied to clinical reasoning. Bordage and Zacks[19] used many of the methods of Rosch to demonstrate that the same kind of graded structure that distinguished the natural categories was present in disease categories. They found, for example, that diabetes was a much more prototypical endocrine disease than Hashimoto's disease or hyperthyroidism. It was volunteered more often by practitioners asked to name as endocrine disease, recognized more accurately and quickly, and so on.

These studies lead to two conclusions: first, there is evidence to substantiate our musings at the beginning of this talk that the concept of disease is a continuum, not a category. Second, the identification of conceptual prototypes such as diabetes, carrot, and robin, which transcend different cultures, argues for an external "nurture" basis for concepts—even high-level concepts such as disease.

Prototype theory, in its methods, seeks evidence for cultural or even transcultural norms for categories. In the extreme, prototype theory might be viewed as empirical evidence for a position that concepts and categories are derived entirely from universals in the environment, a position more extremely nurture-oriented than any we have considered except the positions of Locke and Hume.

Another psychological theory of concept formation, exemplar theory, while still holding to the implicit view that the concepts we learn reflect an external reality, is much more modest about the universality of such concepts. In this perspective, we are able to identify a member of a class or a concept, not because of any internal rules or because the sum of our experience has created prototypes of the class that are available for analysis and introspection, but because we have, in any category (dogs, chairs, diseases, sports cars), an innumerable number of instances of the category (my dog, Rover, Lassie, etc.). When we are faced with a categorization task, a first line of defense is a search through memory for similar examples of the class, and then, if we find an example that is sufficiently similar, we assume the new beast is also a dog. This description makes the process sound far more deliberate and available for introspection than the evidence suggests. Instead, if we inquire why a person decided that the new beast was a golden retriever, the new car was an Audi, or the skin lesion was actinic keratosis, the modal response would be "Because it looks like a golden retriever," or an Audi or actinic keratosis. Further justification may be forthcoming but it sounds suspiciously post hoc. This process is in fact unlikely to be available for conscious introspection.

I and some colleagues have done a series of studies in dermatology[20] and cardiology[21] in which we have found evidence for this mode of processing. As one example,[20] in a series of experiments we gave subjects (residents) practice with a set of dermatology slides covering 11 conditions, then subsequently tested them with a new set of slides. The slides were carefully chosen. Each was drawn from

a quartet of slides containing two typical slides that strongly resembled each other, and two atypical slides that resembled each other. Each subject was then tested with two other slides of the quartet. We balanced it all off, so that we could look at performances on typical–similar, atypical–similar, typical–different, and atypical–different slides. Thus we deliberately compared typicality (a property of the number of features and prototype theory) with similarity (a characteristic of exemplar-based reasoning). The results showed effects of both similarity and typicality. With immediate testing, similarity resulted in a gain of accuracy of about 50%, typicality a gain of about 12%. After ten days' delay, slides that were similar to those in the initial learning series were diagnosed about 25% more accurately, and typical slides were diagnosed about 25% more accurately.

We have continued to explore these phenomena. One concern is that it will work only with visually rich materials, where similarity is highly perceptual. Hatala[21] conducted a study with ECG interpretation, which, while still visual, is replete with quantitative rules. In this study, similarity to an ECG in the learning phase was based entirely on a one-line description (e.g., a "54-year-old accountant" and a "middle-aged banker" versus "an "80-year-old widow"). To demonstrate the effect, the match was to an ECG that was visually similar, but from an incorrect and confusable category (e.g., left bundle-branch block and anterior MI). When the description was matched, accuracy was 23%; when it was unmatched, it was 46%, and of course more residents who saw the matching description fell for the incorrect diagnosis. Further, it would seem that the process must have occurred without awareness. If they had known they were matching on the age and occupation, they would not have done it, since a moment's introspection reveals that this is irrational.

Both of these psychological theories—prototypes and instances—derive from a nurture view of concepts, namely that the concepts we learn are derived from our experiences. In fact, the exemplar models show precisely how specific experiences are available and used in subsequent judgments of category membership. However, as always, there is another side to the story. Psychology has been equally successful at deriving evidence to support the nature view, that what we see is influenced by our own minds. Admittedly, this is not a pure nature view, as we shall see, since the way our perceptions of the external world are biased derives itself from our experience with the world.

Cognitive psychology had its origins in an information-processing model based on the metaphor that the mind is like a computer. However, there was rapid accumulation of evidence showing just how un-computer–like humans are. One simple yet fundamental example is in information retrieval. The answers to questions such as "When did Columbus discover America?" and "What is the capital of Arkansas?" are available almost as soon as you hear the question inflection. Second, if asked about Albania, not Arkansas, you would know that you didn't know almost as rapidly. Contrast that with a search of the Web. Although the computer processes information at least a million times faster than does the mind, retrieval will inevitably take much longer. Further, it will take the computer longer still to decide that it doesn't know, since it will have to search every corner of its memory before it gives up. It is difficult to envision what kind of memory architecture humans must have to do this job, but it must be very different from the computer's RAM.

One model of memory that accommodates these observations is called human associative memory. The model emerged from studies of reading coupled with a phenomenon called the word-superiority effect,[22] which has relevance, surprisingly, to clinical reasoning as well as to many other domains. Imagine that I flash a four-letter word on the computer screen for a few milliseconds and ask you to identify the fourth letter. The phenomenon is this: when the fourth letter occurs in a real word such as "rink" or a pseudo-word such as "bink," the "k" is recognized faster and more accurately than when it occurs in a non-word such as "nrik." While this seems perfectly plausible, it says some fundamental things about the nature of memory. That is, even at the perceptual level of recognizing individual letters, a process that must occur in milliseconds and without conscious introspection, identification is facilitated by memory of much higher-level concepts, the words themselves. This seems to illustrate beautifully the interactive nature of perception, showing that what we see can be influenced by what we expect to see.

The observations of the word-superiority effect were modelled by McLelland and Rumelhart[23] using a "connectionist" or parallel distributed processing (PDP) model, with multiple layers of nodes between input and output corresponding to letter elements, letters, and words, with links among nodes at all layers. Unlike expert systems or Bayesian models, these connectionist models had no preprogrammed rules: rather, they "learned" from experience, gradually building up strength among certain links connecting nodes.

Parallel distributed processing models have been continually refined (and renamed—they are now more commonly known as "neural networks"), and have found application in many settings, including clinical diagnosis, where they appear to be more effective diagnosis machines than the traditional expert systems. However, for present purposes, these applications are less important than the observation that the models have commonality with psychological views of concept formation, based on learning from examples. And as we shall see, the new name is not simply good public relations—neural networks bear a striking resemblance to models emerging from neuroscience.

I and my colleagues have taken the phenomenon that recognition depends, in part, on available concepts in memory into the clinical reasoning lab. In a series of studies in dermatology, radiology, and electrocardiography, we biased the subjects by providing a brief history suggestive of a particular diagnosis, then showed them a visual stimulus—an ECG, a slide of a skin lesion, or a head-and-shoulders picture. We have consistently found that the bias influences not only the differential diagnosis (which might be viewed as perfectly rational), but also the feature calls. Moreover, in a recent study using textbook examples of physical signs,[24] we showed that it was not simply a case that the history increased vigilance for that particular sign, and therefore the likelihood of detection. Rather, an incorrect history led students to misinterpret one sign as another—the inflamed parotid glands of mumps became the moon-shaped face of Cushing's disease, and the moon-shaped face of Cushing's became periorbital edema when linked with a history of nephrotic syndrome.

This phenomenon, that prior higher-level information either provided to the subject or available from memory can influence basic perceptual processes, has been demonstrated at all levels of expertise, from first-year students to cardiologists, so it is not simply a naive bias that can be erased with experience. LeBlanc's follow-up studies of strategies to "de-bias" subjects, under way in our lab, have shown that even fairly draconian measures are only partially successful; a finding that is not surprising since perceptual processes are not available to conscious introspection.

These findings, both in cognition of perception and in clinical reasoning, challenge a commonly held view that experts use "forward reasoning"; that is; they begin with the facts of the case and reason inductively to a logical conclusion, a view championed by Groen and Patel.[25] Their findings were derived from verbal introspections or written summaries, after the subjects had had time to read and reflect on the clinical case. It is my present view that the work on top–down processing, both in reading and reasoning, shows that deductive processes from hypothesized solutions are already occurring long before the case is in full view, and that the apparent induction of the expert simply reflects a coherent story told post hoc. One study done by Eva[26] substantiates this view. He had subjects read mystery stories, than recount their solutions. Half told their solutions "online" as they were reading; the other half,

as a summary after. On three measures, the latter group looked as if they were doing substantially more forward reasoning. However, the manipulation took place *after* the reasoning was over.

What is a Concept? Lessons from Neuroscience

Finally, conspicuous in its absence from the discussion to date is the role of neuroscience in our understanding of concepts. I have described how cognitive psychology has provided examples of phenomena that help us to understand some aspects of clinical reasoning. Theories of concept formation and perception are a useful heuristic for testing apart aspects of clinical reasoning. But the skeptical reader could be forgiven for remarking that these theories seem more like useful demonstrations and analogies than real explanations, in a scientific sense.

Let me then venture into what is for me the largely uncharted territory of neuroscience. In doing so, I am moving closer to the more traditional interpretation of the nature–nurture debate than the way I originally framed it. That is, we now seek evidence from neuroscience that the brain and its structures (nature) are responsive to, and modified by, the environment (nurture). Further, just as basic science provides a framework for understanding disease, neuroscience may provide a framework for understanding the process of concept formation and clinical reasoning.

To advance the neuroscience argument, we need to discover evidence that categories "out there" can be localized to specific brain activities. Perhaps the most accessible argument about the impact of specific experiences on brain anatomy and brain development emerge from the phenomenon of plasticity—the discovery that there are critical periods in the development of the brain during which input from the environment is required in order for specific facilities to develop. The phenomenon is ubiquitous. Here are some examples:

- Children who have congenital cataracts must have them surgically removed before age 10, or they will be unable to recognize shape and pattern, although they will be able to learn colors. This was hypothesized to arise because of abnormal development in the visual cortex. Very recent research with newborns has extended this understanding further. Maurer studied children less than 9 months old who had had cataracts removed immediately following the surgery. Immediately after surgery, their vision was like a newborn's—about 1/40 the acuity of an adult's. But after only one hour of visual input, their acuity had improved to the level of a one-month infant. To quote the researcher: "It's using the eyes and having the experience of seeing that's driving the normal experience of vision after birth. . . . the brain was wired to be ready to receive visual images . . . but it's got to have the input in order to do the learning." [27]
- Animal experiments showed kittens raised in an environment that only allowed horizontal or vertical orientations never learned the other. Hubel and Wiesel[28] then showed that these selective deprivations are identifiable in the development of specific cells in the visual cortex. They went on to show that the brain development was incredibly specific, so that a single day of exposure at day 28 was sufficient to establish the orientation. Other researchers have gone on to establish that plasticity is associated with the presence of specific proteins.

The phenomenon of plasticity is direct evidence of an interaction between brain structures and the environment, and provides an explanation for the philosophical dilemma. Of course, such experiments do not provide direct evidence that higher-order concepts such as temperature, unemployment, love, or for that matter, tables, are associated with specific local changes. The next step is to move from the construction of perceptual maps of the environment to conceptual maps in different areas of the brain. This may not be as large a leap as it sounds; after all, the mechanism that enables us to recognize Aunt Sally must involve links among the more primitive operators that isolate color, shape, and orientation. Thus, we move from brain mappings corresponding to perceptual inputs, which, as we have seen, develop and specialize as a consequence of interactions with the environment at highly specific developmental intervals, to mappings corresponding to the relations among these elements—a "mapping of types of maps," according to Edelman.[29] This remains a theory thus far, the theory of "neuronal group selection." I cannot pretend to be more than an intrigued observer, but it would seem that the evidence at hand regarding neural plasticity provides plausible mechanisms for such a neural correlate of concept formation. Indeed, as I discussed earlier, although neural networks were devised as a simulation device to test a model of concept learning involving parallel and distributed activation, there is a striking correspondence between the nodes and connections of neural networks and the proposed model of neuronal group selection.

Conclusions

This review was intended to accomplish no more than to place the current debates around clinical reasoning in a larger context. There is, in all this, a Michigan State University (MSU) connection. The small research program focusing on clinical reasoning was begun by Elstein and Shulman at MSU in the early 1970s. The McMaster group joined the fray soon after, with me as their hired hand. But soon after this first cycle of studies was completed, there was a strong divergence in the field. Elstein moved his interest to normative approaches such as decision analysis, assuming that clinicians were suboptimal decision makers who could be made more optimal with training. Others who followed, including Patel and Groen, while disagreeing on the details, retained a strongly rationalist perspective. On the other side, Bordage pursued studies in prototype theory, and I began a research program around exemplar models. It is only recently, with the study leading to this review, that I began to appreciate the historical origins of this divergence.

The exciting conclusion from this review is that there appears to be a convergence among the three disciplines—philosophy, psychology and neuroscience—pointing to the reconciliation of these positions. While the constructs, the capacities for identifying regularities appear innate, these abilities are directly responsive to the environment, so that each individual's concepts will be both communal and idiosyncratic. Moreover, this synthesis has some practical implications (believe it or not!). It appears to me that these thinkers are urging us to a reconciliation in our own field—expertise in clinical reasoning is neither mastery of analytical rules nor accumulation of experience, it is both. And the role of experience with individual examples in refining the concepts is critical. Moreover, the philosophical work and the demonstrations of optical illusions show us that the external environment is not delivered to the senses intact, but is filtered through the prisms of prior experience. These are important lessons for instruction in clinical reasoning.

The sum of these findings describes a model of clinical reasoning very different from the algorithmic processes used by the computer (except when, using neural networks, the computer mirrors the mind). An evident implication is that there is little to be gained in demonstrating that humans are suboptimal Bayesians or algorithm-appliers; they are suboptimal because they are using a substantially different basis for computation. While, on the one hand, this provides a strong rationale for computerized decision-support systems, the cautionary note that pervades this review is that the support system cannot intervene after the data are collected, since the data are themselves subject to interpretation in light of mental models.

The Jack Maatsch Memorial Presentation was sponsored by the Office of Medical Education Research and Development, Michigan State University, and presented at the annual AAMC-RIME meeting, October 27, 1999.

Correspondence: Dr. Geoffrey Norman, Department of Clinical Epidemiology and Biostatistics, Health Sciences Centre 2C14, McMaster University, 1200 Main Street West, Hamilton, Ontario L8N 3Z5, Canada; e-mail: ⟨norman@mcmaster.ca⟩.

References

1. Elstein AS. Clinical problem solving and decision psychology: comment on "the epistemology of clinical reasoning." Acad Med. 2000;75(10 suppl).S134–S136.
2. Shulman LS, Elstein AS, Sprafka SA. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press, 1978.
3. Ericsson KA, Smith J. Toward a General Theory of Expertise: Prospects and Limits. Cambridge, U.K.: Cambridge University Press, 1991.
4. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Acad Med. 1990;65:611–21.
5. McGaghie WC, Boerger RL, McCrimmon DR, Ravitch MM. Learning pulmonary physiology: comparison of student and faculty knowledge structures. Acad Med. 1996;71:S13–S15.
6. Papa FJ, Elieson B. Diagnostic accuracy as a function of case prototypicality. Acad Med. 1993;68(10 suppl):S58–S60.
7. Patel VL, Groen GJ, Frederiksen CH. Differences between medical students and doctors in memory for clinical cases. Med Educ. 1986;20:3–9.
8. Bordage G. Elaborated knowledge: a key to successful diagnostic thinking. Acad Med. 1994;69:883–5.
9. Norman GR, Brooks LR, Allen SW. Role of specific similarity in a medical diagnostic task. J Exp Psychol Gen. 1991;120:278–87.
10. Hundert EM. Lessons from an Optical Illusion: On Nature and Nurture, Knowledge and Values. Cambridge, MA: Harvard University Press, 1995.
11. Chalmers AE. What is This Thing Called Science? 2nd ed. St. Lucia, Australia: University of Queensland Press, 1999.
12. Campbell EJ, Scadding JG, Roberts RS. The concept of disease. BMJ. 1979;6193:757–62.
13. Chapman LJ, Chapman JP. Illusory correlations as an obstacle to the use of valid psychodiagnostic signs. J Abnorm Psychol. 1969;74:271–80.
14. Boshuizen HPA, Schmidt HG. The role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. Cogn Sci. 1992;16:153–84.
15. Patel VL, Groen GJ, Scott HM. Biomedical knowledge in explanations of clinical problems by medical students. Med Educ. 1988;22:398–406.
16. Gregory RL. Eye and Brain: The Psychology of Seeing. London, U.K.: Weidenfield and Nicolson, 1966.
17. Deregowsky JB. Illusion and culture. In: Gregory RL, Gombrich EH (eds). Illusion in Nature and Art. New York: Scribner, 1974.
18. Rosch E. Natural categories. Cogn Psychol. 1973;4:328–50.
19. Bordage G, Zacks R. The structure of medical knowledge in the memories of medical students and general practitioners: categories and prototypes. Med Educ. 1984;18:406–16.
20. Regehr G, Cline J, Norman GR, Brooks L. Effect of processing strategy on diagnostic skill in dermatology. Acad Med. 1994;69(10 suppl):S34–S36.
21. Hatala R, Norman GR, Brooks LR. Influence of a single example upon subsequent electrocardiogram interpretation. Teach Learn Med. 1999;11:110–7.
22. Reicher GM. Perceptual recognition as a function of meaningfulness of stimulus materials. J Exp Psychol. 1969;81:274–80.
23. McLelland JL, Rumelhart DE. An interactive activation model of context effect in letter perception. Psychol Rev. 1989;88:375–407.
24. Brooks LR, LeBlanc VR, Normal GR. On the difficulty of noticing obvious features in patient appearance. Psychol Sci. 2000;11:112–7.
25. Patel VL, Groen GJ. Knowledge based solution strategies in medical reasoning. Cogn Sci. 1986;10:91–116.
26. Eva KW, Norman GR. Is thinking aloud equivalent to post hoc explaining. Presented at the 1998 Research in Medical Education meeting, New Orleans, LA. 1998.
27. Maurer D, Lewis TL, Brent HP, Levin AV. Rapid improvement in the acuity of infants after visual input. Science. 1999;286:108–10.
28. Weisel TN. The postnatal development of the visual cortex and the influence of the environment. Bioscience Reports. 1982;2:351–77.
29. Edelman GM. Mountcastle VB. The Mindful Brain. Cambridge, MA: MIT Press, 1978.

144

1999 JACK MAATSCH MEMORIAL PRESENTATION—RESPONSE

Clinical Problem Solving and Decision Psychology: Comment on "The Epistemology of Clinical Reasoning"

ARTHUR S. ELSTEIN

Geoff Norman has presented an extremely rich and stimulating paper that surveys many important themes. In my response to his article,[1] I shall not comment on the connections he seeks between psychology, philosophy, and neuroscience, because this attempted synthesis is well beyond my area of expertise. Instead, my discussion focuses on two other issues: the status of research on the psychology of clinical problem solving, and the connections between this research and decision psychology, the framework in which I have worked for the last 20 years. Then I consider the implications of this work for improving the quality of health care decisions.

Status of Research on the Psychology of Clinical Problem Solving

Several schemes have been put forth to explain how diagnostic reasoning is accomplished, including diagnostic categorization by instance-based recognition,[2] prototypes,[3,4] propositional networks,[5,6] forward reasoning or pattern matching,[7] and generating competing hypotheses.[8] Evidence supporting each of these models is available in the literature. How can this be? Norman argues that no single representation of the process or of the organization of knowledge accounts for all of the phenomena investigators have encountered. Each account is correct sometimes, because individuals adapt their strategies to the demands of the task, including the demands of the experimenter. This implies that experiments designed to test particular hypotheses have also, in some sense, been designed to validate the hypotheses or beliefs of the investigators.

Norman and I agree that problem solvers are adaptive creatures, and we must be careful about concluding that any one account of their behavior will explain all phenomena. He and his collaborator, Henk Schmidt, put it well: "There is more than one way to solve a problem."[9] Viewing problem solvers as adaptive thinkers trying to cope with complexity does not attribute malicious intent either to investigators or to research subjects. On the contrary, it harks back to Newell and Simon,[10] who argued that because of the limitations of working memory, complex tasks are represented in simplified problem spaces, and that consequently understanding problem solving is significantly advanced by understanding that cognitive representation. Their view was quite radical for its time, for the concept of a problem space really committed us to the study of what we now call problem representations or mental models.

Different mental models might be employed by different subjects, or the choice might depend on the task. It follows that a hierarchical organization of medical knowledge, with general concepts at the top and specific instances at the bottom, is a plausible representation and is partially correct. So are propositional networks, with their nodes and connections, symptom-by-disease matrices, and semantic networks. In most studies employing each of these frameworks, the model finds reasonable support in the data. Norman argues that this fit occurs because the subjects, whether medical students, residents, or more experienced physicians, figure out how to adapt to the demands of the task, and these demands usually ask them to behave in ways that provide evidence for the models. This view owes much to Rosenthal's research on demand characteristics.[11] Within the domain of cognitive studies of medical reasoning, I am not aware of studies that test the fits of different cognitive models to the same set of data, so we do not know which would fit the data best or how often each model is used. Studies to test competing models can and should be designed.

Several prominent investigators in the field of medical cognition have used verbal reports of subjects thinking aloud either while solving a diagnostic problem or retrospectively to construct representations of the problem-solving process. Norman notes that "propositional networks are disturbingly idiosyncratic and not apparently reproducible."[1] I cannot entirely endorse his view that "all of these concept architectures are produced on the fly at retrieval, in order to satisfy the expectations of the researcher."[1] It is at least plausible that these "architectures," like other blueprints, are plans for a constructive process: if one follows a blueprint and a house or office building results, we should not be surprised. The plan was designed to lead to that output.

Still, his caution is warranted. We should not unhesitatingly embrace verbal reports as *the* solution to the problem of elucidating cognitive processes. Too much cognitive processing goes on beneath the level of verbal report. And we agree that, to the extent that subjects adapt to the demands of the experimenter, they are likely to tell us what they think we want to hear. These objections imply that research that relies on verbal reports for basic data is not as likely to lead to "truth" as we would like to believe, and that we should move away from thinking-aloud methods back to traditional experimental psychology: the researcher should observe the relationship between the stimulus and the subject's response, and ignore or distrust verbalizations about the task. The subject's response may be verbal, such as a diagnosis or a probability estimate, but a scientific explanation of the thought process should not be based on responses to such questions as "How did you know that?" or "Why do you think this is so?" If these questions are used, we should treat the explanations and justifications as data, not as true accounts of the operations of the subjects' minds.

Both Norman and I have taken these cautionary thoughts to heart over the years. Consequently, we have moved away from thinking-aloud accounts as the primary data source and toward more traditional experimental methods (for examples, see references 12 and 13). We have done this despite knowing that experimental studies will be criticized by clinicians on the grounds that they lack clinical verisimilitude and may not generalize to real clinical settings. A thoughtful clinician will surely ask this question about our work: "Even if I concede that physicians behave as you have shown in this experimental setting, what reason is there to believe that they would behave similarly when dealing with real patients?" Anticipating this question, Norman and his colleagues have worked extensively with visual stimuli, such as radiographs and EGG tracings, that are unquestionably part of the real clinical world.[14,15] But this strategy begs the question, "Do the results apply to non-visual stimuli, such as are obtained in taking a good history?" My colleagues and I have done some research using case

vignettes[16–18] that do not use thinking aloud to study clinical reasoning. One objection raised to our findings in those studies relates to motivational factors: clinicians are not motivated to do their best with hypothetical cases and would do "better" with real patients. I think it unlikely that clinical problem solving will be better in complex environments, with many distractions, than in simplified laboratory settings, but I concede that, just as with pharmaceutical research, laboratory findings should be verified in the "real world." Nobody ever said that doing good research would be easy.

Decision Psychology

Norman noted that my own research program moved in a different direction after 1980, from a focus on clinical "problem solving" to "decision making." What is the difference? For over two decades, much of the research on the psychology of decision making has been dominated by statistical decision theory, a model of idealized rationality under uncertainty. Behavioral decision research has concentrated on identifying systematic departures from this model, and these departures are viewed as "errors." The research has shown that while decision theory may be an account of ideal rationality, it is not a description of how people actually make judgments and choices under uncertainty. In short, limited rationality has its impact on both decision making and problem solving. The psychological processes that produce these errors are called "heuristics and biases." Indeed, the entire line of research has come to be identified by this term.[19,20]

Norman argues that there is not much point in identifying cognitive heuristics and biases that violate the rules of statistical decision theory, since people are not trying to reach conclusions using these principles. To quote: "An evident implication is that there is little to be gained in demonstrating that humans are suboptimal Bayesians or algorithm-appliers; they are suboptimal because they are using a substantially different basis for computation."[1]

In my judgment, Norman has misunderstood the research agenda of decision psychology and its implications for medical education. The study of clinical diagnostic reasoning from the problem-solving point of view implies one thinks of diagnosis as categorization. The research questions then center around issues such as, "What categories does the problem solver know? What features justify placing the case in one category or another?" These are questions about the knowledge base and feature recognition and interpretation. From the decision-making standpoint, clinical diagnosis is opinion revision with imperfect information, and treatment choice is about how best to balance benefits and harms. Risk and uncertainty are everywhere. The aims of the research are to identify the processes people use in making complex judgments and choices under these conditions, and to ask whether their behaviors are consistent with Bayes' theorem (for diagnostic reasoning) and maximizing expected utility (for treatment choices). If behavior is not consistent with these principles, and if we find these principles sensible and appealing, we might well wonder what kind of educational program could be developed to improve our decision making. Therefore, there is just as much point to studying cognitive heuristics and biases as there is to studying the roles of instances and prototypes in categorization. Indeed, the role of instances in categorization can be seen as a special case of base-rate neglect or of treating irrelevant data as strong evidence: in reality, some of the cues associated with the instance have likelihood ratios close to 1.0 (the decision-theoretic definition of irrelevant), but are treated as if they are meaningful, say >10.0. Using two very different theoretical frameworks, both of us have thrown some light on how the mind works, and we have shown that human inference can be improved upon. To improve clinical decision making, it seems to me that decision theory is at least as promising as the study of categorization processes. I still think that a general strategy applicable to a wide range of clinical situations would be very useful in helping people to think straight. Norman referred to the finding of content specificity, discovered in my early research in this area.[8] Given this fact, the need for a general approach to sound thinking is even greater than we had previously suspected.

What is the evidence that clinicians at times need help in thinking about complex problems? Two related bodies of evidence, from cognitive psychology and from health services research, support this claim. From cognitive psychology, we have a series of lessons and findings about limited rationality. Health services researchers have provided a growing body of literature on practice variation (for example, see references 21 and 22), which has repeatedly shown that something besides hard science is involved in many medical decisions, both diagnostic and therapeutic, and that these variations are not necessarily rational responses to differences between patients.

How Can We Improve the Quality of Clinical Practice?

Interestingly, in the past 20 years, two related decision technologies have arisen that deal precisely with these issues: evidence-based medicine (EBM) and decision analysis (DA). Both offer to the medical community ways of quantifying the evidence, dealing with uncertainty and error in the evidence, and trying to systematically weigh risks and benefits of alternative treatment strategies. The rapid dissemination of these principles may be attributed in part to the diligence and enthusiasm of their devotees, but it cannot be entirely explained by their efforts. The zeitgeist or cultural climate had to be ready. In my view, psychological research on problem solving and decision making has contributed to these developments by showing that expert clinical judgment was not as expert as we had believed it to be, that knowledge transfer was more limited than we had hoped it would be, and that judgmental errors were neither limited to medical students nor eradicated by experience. EBM and DA offer approaches for dealing with these problems, and that is why they are making headway in clinical medicine. Clinical practice guidelines, which are intended to improve the overall quality of care, are another, related, approach to these issues, and the problems encountered in their dissemination and implementation have been widely discussed.[23–26]

The reactions to these approaches suggest that the tension between theory and practice will remain. All theories and models are simplifications of reality. They abstract particular features in order to provide a reasonably coherent account of how things work and to guide action. That is precisely why they are useful. Models are not reality, however, and theory is not practice. Consequently, physicians often mistrust the adequacy of scientific accounts or guidelines based on evidence, despite the necessity of relying upon them. Because general principles will never be able to account for all concerns in clinical cases, there will always be room for judgment, applying general principles on a case-by-case basis.

Encomium: Let Us Now Praise . . .

Geoff left his comments about the connection to Michigan State University (MSU) and its College of Human Medicine for the close of his remarks, and I follow his example. How fortunate that many years ago, Geoff Norman came to MSU and joined our small group of scholars. We were not aware that we were doing classic work that would be argued and discussed and revisited for a generation. Who could possibly have thought that? Yet, if there was ever a golden era of research in medical education, it was there and then. We have made some progress, and we have had a wonderful run. When I think of that medical school and its faculty and students back in the 70s, the wonderful line from Shakespeare's *Henry V* always comes to mind: "We few, we happy few, we band of brothers." How appropriate that in Jack Maatsch's memory we have come together to discuss some issues that concerned him and to celebrate that happy band!

146

The Jack Maatsch Memorial Presentation was sponsored by the Office of Medical Education Research and Development, Michigan State University, and presented at the annual AAMC–RIME meeting, October 27, 1999.

Correspondence: Dr. Arthur Elstein, Department of Medical Education, University of Illinois at Chicago, 808 South Wood Street, Chicago, IL 60612-7309; e-mail: (aelstein@uic.edu).

References

1. Norman GR. The epistemology of clinical reasoning: perspectives from philosophy, psychology, and neuroscience. Acad Med. 2000;75(10 suppl):S127–S133.
2. Brooks LR, Norman GR, Allen SW. Role of specific similarity in a medical diagnostic task. J Exper Psychol Gen. 1991;120:278–87.
3. Bordage G, Zacks R. The structure of medical knowledge in the memories of medical students and general practitioners: categories and prototypes. Med Educ. 1984;18:406–16.
4. Bordage G, Lemieux M. Semantic structures and diagnostic thinking of experts and novices. Acad Med. 1991;66(9 suppl):S70–S72.
5. Patel VL, Evans DA, Groen GJ. Biomedical knowledge and clinical reasoning. In: Evans DA, Patel VL (eds). Cognitive Science in Medicine. Cambridge, MA: MIT Press, 1989.
6. Patel VL, Evans DA, Kaufman DR. A cognitive framework for doctor–patient interaction. In: Evans D, Patel V (eds). Cognitive Science in Medicine. Cambridge, MA: MIT Press, 1989:257–312.
7. Patel VL, Groen G. Knowledge-based solution strategies in medical reasoning. Cogn Sci. 1986;10:91–116.
8. Elstein AS, Shulman LS, Sprafka SA. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press, 1978.
9. Norman GR, Schmidt HG. The psychological bas. of problem-based learning: a review of the evidence. Acad Med. 1992;67:557–65.
10. Newell A, Simon HA. Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall, 1972.
11. Rosenthal R. Experimenter Effects in Behavioral Research. New York: Appleton-Century-Crofts, 1966.
12. Hatala R, Norman GR, Brooks LR. Influence of a single example on subsequent electrocardiogram interpretation. Teach Learn Med. 1999;11:110–7.
13. Elstein AS, Christensen C, Cottrell JJ, Polson A, Ng M. Effects of prognosis, perceived benefit and decision style upon decision making in critical care. Crit Care Med. 1999;27:58–65.
14. Norman GR, Brooks LR, Coblentz CL, Babcook CJ. The correlation of feature identification and category judgments in diagnostic radiology. Mem Cogn. 1992; 20:344–55.
15. Regehr G, Cline J, Norman GR, Brooks L. Effects of processing strategy on diagnostic skill in dermatology. Acad Med. 1994;69:S34–S36.
16. Christensen C, Heckerling PS, Mackesy-Amiti ME, Bernstein LM, Elstein AS. Pervasiveness of framing effects among physicians and medical students. J Behav Decis Making. 1995;8:169–80.
17. Bergus G, Chapman GB, Gjerde C, Elstein AS. Clinical reasoning about new symptoms in the face of pre-existing disease: sources of error and order effects. J Fam Pract. 1995;27:314–20.
18. Chapman GB, Bergus GR, Elstein AS. Order of information affects clinical judgment. J Behav Decis Making. 1996;9:201–11.
19. Elstein AS. Heuristics and biases: selected errors in clinical reasoning. Acad Med. 1999;74:791–4.
20. Kahneman D, Slovic P, Tversky A (eds). Judgment Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press, 1982.
21. Vehvilainen AT, Kumpusalo EA, Takala JK. They call it stormy Monday—reasons for referral from primary to secondary care according to the days of the week. Br J Gen Pract. 1999;49:909–11.
22. Lim LL, Heller RF, O'Connell RL, D'Este K. Stated and actual management of acute myocardial infarction among different specialties. Med J Aust. 2000;172: 208–12.
23. Greco PJ, Eisenberg JM. Changing physicians' practices. N Engl J Med. 1993;329: 1271–4.
24. Asch DA, Hershey JC. Why some health policies don't make sense at the bedside. Ann Intern Med. 1995;122:846–50.
25. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA. 1999;282:1458–65.
26. Poses RM. One size does not fit all: questions to answer before intervening to change physician behavior. Joint Comm J Qual Improvement. 1999;25:486–95.

147

1999 INVITED ADDRESS

The Marvelous Medical Education Machine or How Medical Education Can Be Unstuck in Time

CHARLES P. FRIEDMAN

The jumping-off point for this paper is actually the second part of its compound title. The concept of becoming "unstuck" in time stems from the initial line of Kurt Vonnegut's popular novel *Slaughterhouse Five* [Vonnegut 1969]:

> Listen: Billy Pilgrim has come unstuck in time.

In this paper I will actually argue that medical education has become "stuck," not only in time but also in space and content. It has become stuck in time because events considered to be educational largely occur through interactions that require the learners and the faculty to be simultaneously participating in these interactions. It has become stuck in space because its mechanisms of delivery are largely bound to a specific physical location, the academic medical center with its classrooms and associated health care delivery venues. It has become stuck in content because the topics that are the focus of educational interactions are insufficiently under the control of the students, and the teachers. Increasingly, there is no reason for any of these requirements to be imposed on the educational process. Moreover, medical education remains stuck in an era when much of the rest of human enterprise is becoming unstuck, the result of a sweeping set of cultural changes made possible by information technology and primarily by the phenomenal proliferation of the global Internet [Drucker 1999].

I will further argue in this paper that medical education can gradually be "unstuck" in space, time, and content through appropriate use of emerging technology, with emphasis on simulation methods that have become widespread in the use of training pilots and professionals in other disciplines. Modern flight simulators have become so sophisticated that experienced pilots being certified to fly a new aircraft might have a load of passengers in the back the first time they actually fly the plane [Dawson and Kaufman 1998]. While there will always be a pilot experienced flying this aircraft alongside the neophyte in the cockpit, this practice clearly testifies to the educational power of simulations. Recently, the U.S. Navy adopted the inexpensive Microsoft "Flight Simulator"® program as standard training for its new pilots, after a trainee who practiced extensively on this program recorded the best performance ever on an initial training flight [Brewin 2000].

The "marvelous medical education machine," as the concept will be developed in this paper, is the complete simulator for medical education, analogous to the best of contemporary flight simulators. But like Vonnegut's novel, the marvelous machine is currently a work of fiction. It does not exist although bits and pieces of it do exist, and these suggest what might be possible in the not-too-distant future. In the sections that follow, I will describe the need for the marvelous machine in greater detail, discuss what it can potentially do when built, expose the internal anatomy of the complete machine, review some of the pieces that exist now and how we might build it from here, and finally discuss some of the key educational research questions that will have to be illuminated along the way. This paper, in its entirety, will argue that building the marvelous machine should be a top priority for medical education nationally and internationally.

Stuck in Space, Time, and Content

To clarify what it means for medical education to be "stuck," it may be useful to consider education as a process with events that exist in three dimensions (Figure 1). The first dimension can be thought of as physical space, the second time, and the third the biomedical topic that is under consideration. Medical education is stuck in all three dimensions, because teachers and learners have little control over these dimensions: where and when the events occur and what topics are addressed. In the basic sciences, for example, lectures and labs occur in a fixed place and at a scheduled time and on a topic that faculty believes the students need to know about—and then they are over. In the clinical sciences, patients (who remain the primary "teaching material" even though this term is seldom used anymore) appear at a fixed location and at a particular time with the problem they happen to have—and then they leave.

This way of doing educational business is so much a part of daily life in an academic medical center that most of us take it for granted; and since our students learn and graduate and become certified as practitioners, it is easy to conclude that there is nothing wrong with being "stuck." But there are profound reasons for concern. First and foremost, education that is stuck routinely ignores much of what is known about teaching and learning in medicine. Studies of clinical reasoning accumulated over more than 20 years point to the "case specificity" of medical expertise, meaning that proficiency generalizes very weakly from disease to disease and, more generally, from one aspect of medicine to another [Elstein et

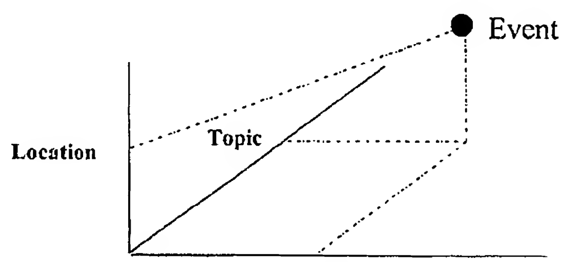

Figure 1. Traditional medical education is "stuck" in the dimensions of space, time, and content.

This article is based on the annual invited address of the same name delivered by the author at the 38th Research in Medical Education Conference during the meeting of the Association of American Medical Colleges, Washington, DC, October 27, 1999. The article, reprinted here with permission of Taylor & Francis Ltd., was first published in *Medical Teacher* (2000;22:496–502) as part of the conference proceedings of the annual meeting of the Association of Medical Education in Europe, Beer Sheva, Israel, August 27–30, 2000. The text of the article is identical to the original version except for minor changes in citation style in the reference list.

148

al. 1978; Schmidt et al. 1990]. As such, the most effective way, and perhaps the only way, of developing proficiency over time is active practice with a wide range of cases and with as many repetitions for each subject/disease area as possible [Issenberg et al. 1999]. In educational environments that are stuck, live patients are the primary source of such practice; yet faculty and students have no control over the patients who walk into the clinic or are admitted to the hospital. Active, appropriate practice under these circumstances can be very difficult to engineer, much less guarantee.

Another problem is the expectations of a coming generation of learners that has increasingly "grown up digital" [Tapscott 1998]. Our students who have experienced increasingly sophisticated video games, and who have spent hours with such excellent simulations as Sim City® and Flight Simulator®, will recognize immediately the potential for similar experiences to enhance their training in medical domains. These learners will intuitively understand that medical education is stuck in space, time, and content. Although they may not use these exact words, they will find being stuck unacceptable. They may articulate this recognition by comparing their medical education experience with their undergraduate experience, wondering why, as the sophistication level of what they are studying is increasing, the sophistication of the technology used to support these studies is decreasing. In the short term, they may accept what they see as antediluvian educational practices, simply because these represent the only pathway to a desired profession, but over time they will demand a different kind of service, the need for which and the practicality of which they see as self-evident. If they cannot get this service from traditional educational institutions, their instincts honed by the Internet culture will lead them to seek it from other sources.

Economic pressures on academic medical centers may drive change as well. The problem of providing appropriate practice for trainees exacerbates as health care economics shortens hospital stays and clinic visits, and trainees necessarily have more limited access to patients. Clinical faculty members at academic medical centers and in community settings may perceive that their productivity is judged much more by patient throughput than student learning. An educational system already limited in its ability to provide an appropriate range of "teaching material" may find itself unable to provide appropriately motivated teachers as well.

If academic medical centers do not systematically recognize the opportunity afforded by information technology to "unstick" the system, others will. Hafferty has warned that, for a variety of reasons, medical education based in academic centers could lose its social mandate by not addressing in the curriculum a widely-recognized set of social needs, and thus become irrelevant to the needs of the modern world [Hafferty 1999]. Similarly, by remaining obstinately stuck in space, time, and content, academic medical centers could lose what may be called their "technical mandate" to educate because the methods being used no longer make sense to trainees and to society as a whole. Simultaneous loss of social and technical mandates will generate alternative approaches to education that could, over time, become the norm. Such alternatives are already becoming evident, for example, in the Open University's plan to offer a curriculum equivalent to the first two years of the medical curriculum in the United Kingdom [Daniel 1999], and possibly through Internet ventures such as "medschool.com"™ [Medschool.com 2000]. Established academic medical centers can choose to be leaders and active partners in these developments, or not.

Some may ask to what extent the technique of standardized or simulated live patients [Ainsworth et al. 1991], which has occupied much of the attention of the medical education research community over the past two decades, offers the capabilities of the marvelous machine. It does, but as a practical matter only to a very limited extent. Standardized patients are expensive and do not offer the economies of scale that, as will be discussed, the marvelous machine so profoundly offers. The largest expense associated with use of standardized patients is the wages they must be paid, and the 20th standardized patient encountered by a student costs almost as much as the first. Standardized patients must be painstakingly trained, and there are significant costs associated with this training that are completely lost once the patient retires from active educational service. And a standardized patient can offer only limited variations on the case he/she was trained to represent. As a trainer for procedures, standardized patients must endure the mistakes of the non-expert. Invasive or risky procedures cannot ethically be performed on them at all. Although they can explain how they feel, standardized patients have no access to what is actually happening inside their bodies, and cannot explain to trainees the consequences of their actions at the organic or cellular levels. Finally, standardized patients cannot easily record what is being done to them by the trainee, so feedback to trainees cannot be related with high precision directly to their actual actions and decisions. So while standardized patients can be enormously valuable sources of practice and tools for assessment, they take medical education only part of the way to where it can and needs to go. They are, for the most part, stuck in space, time, and content.

Potential of the Marvelous Machine

Remember above all that the marvelous machine does not currently exist. As we consider what the future might hold if medical education begins a steady progression toward the development of the marvelous machine, it is useful to visualize an end-point of this progression. I do not envision, ever, the complete elimination of teaching around live patients in the same sense, although some might disagree, that no novice pilot is likely to receive a license without flying a real plane. Nor does this work envision that neurosurgery residents will perform their first operation solo after five years of practice only on a simulator. I do, however, envision a future where medical trainees, and practitioners for their continuing education, spend increasingly large fractions of their time working on computer-based simulators. The reasons for this are examined below. Later sections explore what must be inside such a machine in order for it to do these things.

The marvelous machine is unstuck in the three-dimensional educational space described earlier (see Figure 1) because it can provide tireless practice of medical diagnosis, management, and clinical procedures. It is unstuck in the time dimension because it can be used anytime, for as long as the trainee wishes, and over and over again to provide the kind of meaningful repetition of tasks that is highly desirable. The machine is unstuck in the space dimension because the ideal, fully developed machine can "go" or be accessed anywhere. The Internet can in principle bring the capabilities of the machine to a trainee at home or on campus, anywhere in the world. This capability has enormous implications for the future of medical education as it requires us to think of the medical school not so much as a physical place but as a set of learning resources that can be delivered anywhere [Friedman 1996]. The machine is unstuck in the content dimension because it can address on demand topics and skills of faculty and/or student choice, creating appropriate variants of each case or topic to enable meaningful practice to occur. It can record every element of what happened during a student's work with a case—generating highly specific feedback to the learner on his/her performance and informing student and faculty choices about what further practice each student may need.

A further key feature of the marvelous machine are the fortunate economics of its use. Once developed and programmed, there is minimal marginal cost attaching to its operation. The contrast to standardized patients, who are paid a fixed sum per hour, is particularly striking in this regard.

To understand the potential of the machine from a somewhat

149

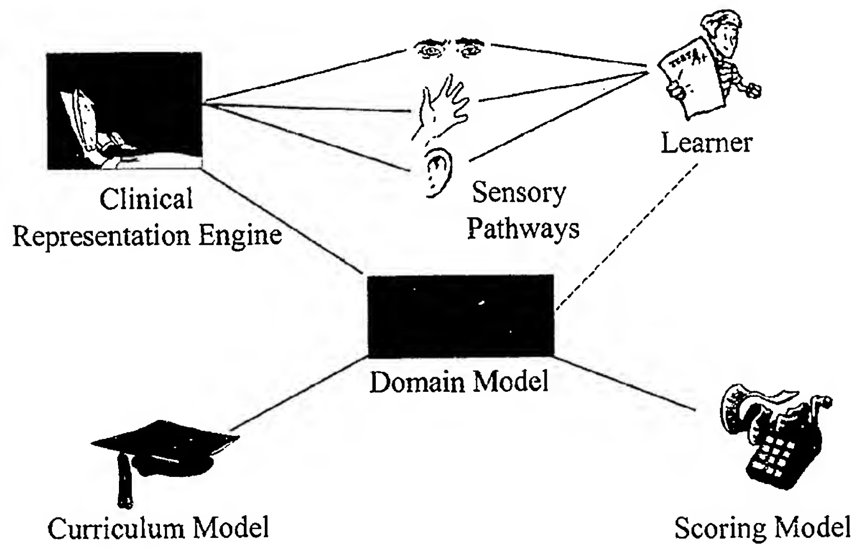

Figure 2. The anatomy of the marvelous machine.

different perspective, consider the potential of such a device to engage learners in "what if" games, which are enormously educational. Students can ask, and get answers from the machine, to the following classes of "what if" questions:

- *What if* I did the procedure again, just a little bit differently? The marvelous machine allows students to tinker in a way that enables them to hone their skills and judgment. A student can, for example, explore the consequences of perhaps giving a slightly stronger dose of a drug to a "patient" whose disease is being simulated by the machine.
- *What if* I did this in a way I know is wrong? Without the machine, it is difficult to experience the consequences of mistakes as a way to learn to manage them. In the real clinical world, mistakes certainly cannot be purposely made, and when they occur occasionally by accident or oversight they are not often not recognized as such until long after their occurrence. With the machine, students can make mistakes on purpose, knowing that they are mistakes, so they can practice managing the consequences, or just to see what happens.
- *What if* I did this 100 times in each of two different ways? One of the most educationally creative ways of using the marvelous machine may be to conduct an "instant clinical trial" by instructing the machine to treat 100 instances of the "patient" one way and another 100 instances a different way. The models built into the mature machine, as will be discussed below, are necessarily and realistically probabilistic and the machine will therefore reflect naturally occurring variability in the way organisms respond to drugs and other external stimuli.
- *What if* biology worked just a bit differently? Used in this way, the machine can connect the basic and clinical sciences. In a fully mature version of the machine, students can be given the capability of changing the parameters of the biological models that drive the simulator. The potential of enhancing their understanding of basic biology is significantly enhanced through the ability to see how organisms would act or react if the basic laws of biology were constructed just a bit differently from the way we believe they are.

Based on this, for now, somewhat abstract conceptualization of the marvelous machine and using our imaginations to conceive what someday the machine will be able to do, consider how medical education must then be undertaken. Medical education would not look and feel at all as it does now. The rationale for "lockstep" learning wherein students proce ' in unison through a relatively rigid curriculum would disappear almost completely, and likely with it would disappear the notion of a four-year curriculum. Indeed, lockstep learning can be seen as an administrative artifact of the lack of a mature marvelous machine. Students could, in principle, begin their predoctoral education whenever they were ready, and authorized, to do so. They would end it when they had proved they had mastered the stated objectives of the curriculum. There might be no need to have students physically on the central campus most, or even some, of the time. Lectures certainly have their place as an educational medium, but the current reliance on lectures as a primary mechanism for conveying information would no longer make sense. Perhaps, with the marvelous machine, we could return to the pre-Flexnerian concept of the part-time student without inheriting the educational inadequacies of the pre-Flexner era. While this paper does not focus on continuing education of physicians and other health professionals, the needs in continuing education are such that the potential effects of the machine on this level of the educational continuum are similarly revolutionary [Barnes 1998].

The Anatomy of the Marvelous Machine

Now to more technical specifics. How are we going to build the machine? What is its anatomy [van Meurs et al. 1997], its necessary component parts?

As illustrated in Figure 2, the marvelous machine can be seen as having five major components, not counting the "learner" without whom the machine would have no purpose. The specific techniques for developing each of these components are beyond the scope of this paper, but a later section discusses the academic disciplines that contribute to each one.

- First and foremost, the machine has a **domain model**, which is a mathematical description of the biological phenomena governing the disease or body sub-system of interest. The domain model computes the state of the patient and the effects of the learner's actions on the state of the patient. The mathematical domain model is what makes it possible for the marvelous machine to generate an endless supply of novel cases and other practice opportunities, and it is what largely sets the marvelous machine apart from traditional simulation environments that use

"scripted" cases. Typically, these domain models have explicit probabilistic features that reflect the natural variability in disease development and response to clinical interventions.

- Next is the **clinical representation engine**. This component is necessary because the output of the domain model is typically a set of numbers that must be translated into clinical observables: statements the patient would make about his/her disease ("I feel tired all the time . . ."), findings that could be appreciated on physical examination ("The patient is cyanotic . . ."), and test results ("Biopsy reveals a tumor . . .).
- The **sensory pathways** component takes the findings and creates portrayals of them that are actually seen, heard, or touched by the learner. This component can be seen as the virtual reality aspect of the machine [Hoffman and Vu 1997; Satava and Jones 1998]. In a mature version of the machine, the learner will see the patient and hear his/her statements; and experience his/her physical condition through sight, touch, and hearing. All of these presentations would change as the patient's condition changed, as directed by the domain model in response to actions taken by the learner and/or a natural evolution of the patient condition. The changes might occur in real time, as would be the case if the learner was using the machine to practice a procedure, or in compressed (simulated) time if the learner was using the machine to practice longitudinal management of a chronic disease.
- The **scoring model** is the basis of providing performance feedback to the learner. Although there are other ways of approaching this problem, the scoring model typically would compute what is the ideal action for the learner to take at any point in the simulation and compute an instantaneous "score" for learner through a metric that compares the ideal performance with what the learner actually did. In some versions of the machine, the knowledge encoded in the domain model can also be harnessed to power the scoring model.
- Lastly, a complete educational application using the machine must have a **curriculum model**. Since the machine's domain model can support learners' practice by constructing cases with specific problems and other characteristics, the curriculum model would represent the set of problems and characteristics on which all learners must have practice, and in which order. For each learner, the curriculum model would maintain records of which aspects of practice had actually occurred.

To illustrate how these components of the machine would interact to generate a comprehensive practice experience, we could follow the machine through one conceptual cycle of operation. This example is a bit simplistic, but illustrative of the concepts.

Ms. Smith, a medical student, is taking a rotation in clinical oncology and indicates to the machine that she wants to practice on a simulated case. The process begins with the **curriculum model** determining that she has not completed her minimum quota of practice on managing metastatic breast cancer. The machine may ask Ms. Smith at that point if she would like some further practice in breast cancer management. After an affirmative response, the **domain model** then generates a case of metastatic breast cancer, represented mathematically, subject to the constraints passed to it by the curriculum model. Because the domain model is inherently probabilistic, many features of the case presented to Ms. Smith are determined by chance and no two cases would be exactly the same. The **clinical representation** engine then converts the initial state of the patient to a set of clinical findings that can be made known to Ms. Smith, should she request them as part of her initial work-up of the patient.

Ms. Smith's work then begins. She is told that the patient is in her "clinic" and takes a history, performs an exam and runs tests on the patient. Only those patient findings actually requested by Ms. Smith would be revealed to her. This is mediated through the **sensory pathways** component of the machine. Ms. Smith would hear the patient's voice responding to questions, see (and, depending on the maturity of the virtual reality component of the machine, perhaps feel) the areas affected by the patient's previous surgery, and see the results of lab tests and imaging studies indicating metastatic disease. Based on Ms. Smith's initial work-up, she then puts the patient on a regimen of chemotherapy.

The **domain model** then computes the effects of the chemotherapy on the course of the patient's disease, mathematically modeling the growth of tumor cells, the reactions of these to the therapy, and any toxicity that may result from the therapy. The **scoring model**, in the meantime, has assigned and recorded a score (or scores) to the actions Ms. Smith has taken.

Assuming that the domain model determines that Ms. Smith's therapeutic regimen would cause toxicity, Ms. Smith would encounter the patient again when that toxicity had developed to the point that the patient would be symptomatic and would return to the clinic. At this later point in simulated time, the domain model will have generated a new set of mathematical parameters describing the patient's updated condition, and will have passed them to the clinical representation engine. The cycle of the machine's operation continues with Ms. Smith having the opportunity to examine the patient again, run more tests, and make decisions to manage the toxicity. Those decisions would be assigned a score, and the simulation would continue until the exercise was completed. Ms. Smith might indicate to the machine that she was finished, or the patient might die or become disease free after a sufficient period for the domain model to conclude a probable cure. It would then be possible for Ms. Smith to initiate a dialog with the scoring model, which would present her score and critique her performance. If Ms. Smith wished, she could run the simulation clock back to a point where her performance was sub-optimal, and play a "what if" game by trying something different and experiencing the consequences of her revised actions.

How the Machine Will Be Built

Is the example above science fiction? Partially, but on balance, not. Indeed, a primative version of the simulator described above (see Figure 3) has been developed through the OncoTCap project at the University of Pittsburgh [Day et al. 1998]. A key innovative element of OncoTCap · the development for many specific areas of oncology of a domain model that is powerful enough to drive a simulation of the type described. Indeed, OncoTCap's domain model allows students to play all the "what if" games described earlier. They can try again, just to see if they can do better; they can do something wrong on purpose, just to see what happens or to practice managing the consequences, they can instruct the domain model to run two types of treatment, each with 100 simulated patients, to run a "clinical trial" to see which method is superior; and they can even change the parameters of the domain model to see what the world would be like if biology worked a bit differently than science currently thinks it does.

Other notable efforts to build elements of the marvelous machine are described below. Still, it is safe to say that building the marvelous medical education machine *is* rocket science. It is much harder than building a flight simulator, in part because of a major difference between aviation and medicine. As Dawson and Kaufman (1998) have observed, in medicine one must manipulate the environment whereas in aviation the goal is to avoid it. The problems of realistically representing clinical findings, and their evolution over time in the same patient, are enormous. When one loads, on top of that, the virtual reality aspects of creating the sensation of actually interacting with the patient through all senses, the full magnitude of the challenges that lie ahead begins to come clear.

So how will the machine be built? First of all, it will be built

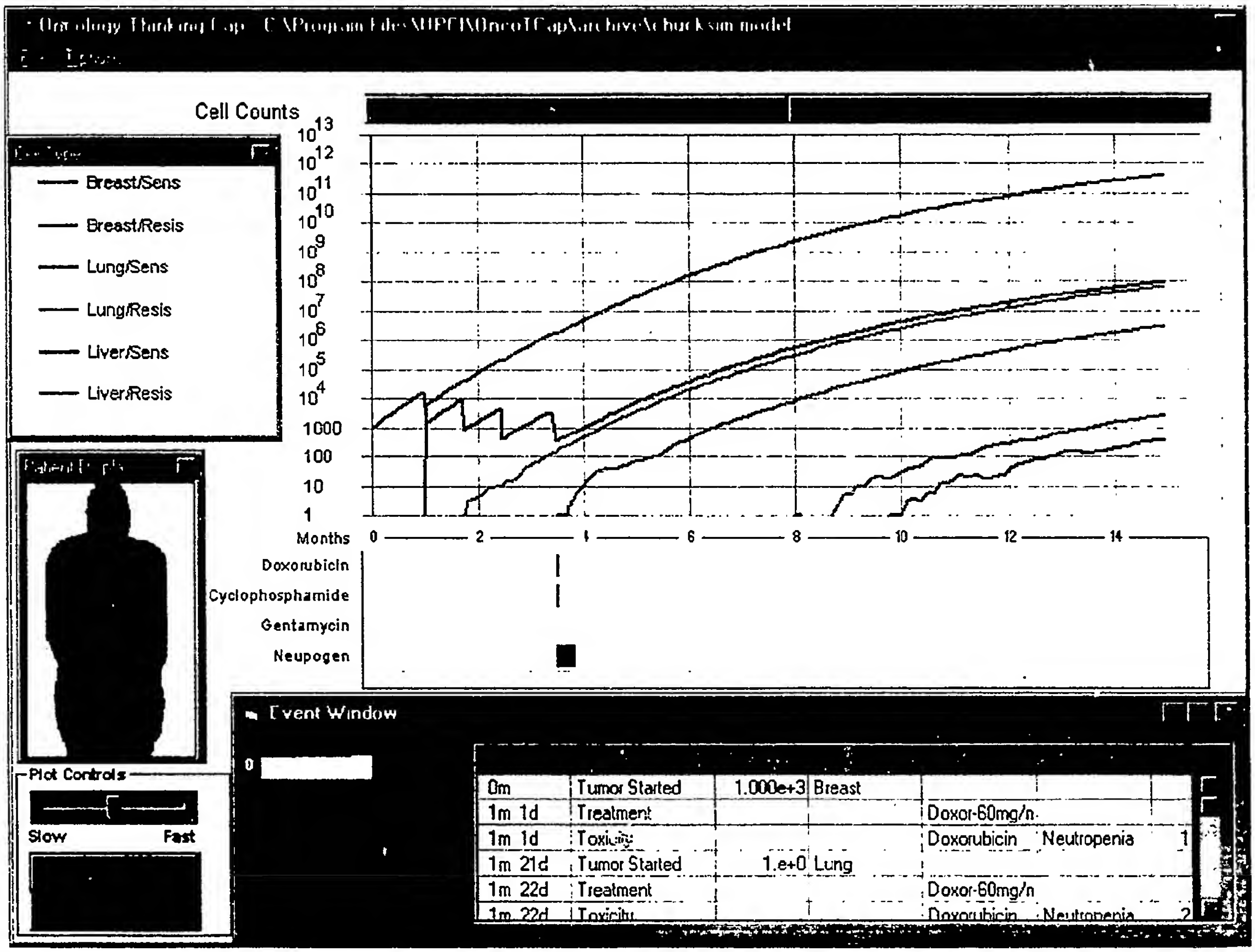

Figure 3. Screen shot of the OncoTCap simulator which has been linked to a clinical presentation engine.

incrementally. Pieces of it already exist; other pieces will arrive in the near future; and in some sense it will never be complete. It will just get better and better over time. Second, it will be built domain-by-domain. The comprehensive unified mathematical model of human biology, the "Maxwell's Equations of biology," probably do not exist and, if they do, they are not likely to be discovered anytime soon. What we are therefore likely to see in near future are cancer simulators, anesthesiology simulators, diabetes simulators, surgical simulators, etc. These domain-specific simulations will become increasingly sophisticated, and then at some point in the future, the models will become sufficiently powerful that simulations from different domains will begin to merge. Finally, the marvelous machine will be developed through collaborations among clinical domain experts and scientists in various disciplines. The clinical representation and sensory pathway components are problems that fall to computer scientists and engineers; domain modeling is work for computational biologists and researchers in artificial intelligence; the scoring and curriculum models are the purview of psychometricians and decision analysts.

Collaborative efforts to develop components of the marvelous machine abound. Many collaborations, some of which have created mature products, have been ongoing for many years. To cite just a few examples, two models of simulators for anesthesiology have reached a high level of development [Norman and Wilkins 1996]. A group in London has developed a prototype marvelous machine for diabetes [Lehmann 1998]. Groups at Stanford and UC–San Diego have taken important strides in developing anatomical simulations that are the basis for building practice on clinical procedures into the marvelous machine [Hoffman and Vu 1997; Dev et al. 1998], as have groups at Mayo and Walter Reed Hospital in the specific area of GI procedures and endoscopy [Robb 1997]. It is important to acknowledge the CBX (Computer Based Exam) project of the National Board of Medical Examiners, which has created a comprehensive simulation environment of the U.S. medical certification process [Clauser et al. 1998] and an effort underway at the American Board of Family Practice to develop simulations driven by mathematical models [Sumner et al. 1998]. Algorithms that would drive a scoring module of the type described above have also been developed [Downs et al. 1997].

So while the marvelous machine as a whole does not exist, it is very safe to say that significant bit and pieces of it do exist and there are substantial reasons to believe that it can and will be built over time.

Conclusion: The Educational Research Challenges

A final piece of the challenge of the marvelous machine is the set of educational research questions that must be addressed if the machine is going to be built properly and its value and place in medical education thoroughly understood. By this I do not mean the

myriad of technical research challenges that will have to be overcome to build the domain models, scoring models, clinical representation engines, and sensory pathways to the learner. As discussed earlier, these fall properly into the research areas of computer science and engineering, computational biology, and other fields.

From an educational research perspective, the key questions map out uncharted territory because of the novelty of what the machine can do. To the extent that the machine represents new technology with the potential to be of benefit, this does not mean that the machine *will* be of benefit. As with any technology, there is potential for it to leave us less well off than we were before. In the end, no matter how well the technology itself functions, the success of the machine will depend on how the machine is used, the educational engineering of the machine into a comprehensive learning environment in which the machine is but one element. As noted earlier, teaching around live patients is not going to go away, no matter how sophisticated the machine becomes over time. Even though the live lecture as an educational medium is completely stuck in space and time and content, the live lecture will likely prove more durable than its most strident critics would have us believe. The proper use and integration of the machine into medical education can be directed profoundly by research that addresses questions such as:

- Relative to the domain model, how "good" do these models have to be in order for them to be ready for use in education. For educational purposes it is perhaps sufficient for the domain model to create and evolve cases that are plausible, but not absolutely correct [Friedman 1995]. But how plausible is plausible enough?
- Relative to the curriculum model, how should an unstuck curriculum be structured? With freedom to learn anywhere, anytime, and on topics of student and/or faculty choice, how much freedom is the right amount of freedom? What should be constrained? To what extent should the domain learning model be one of discovery-oriented? How should more, or perhaps less, freedom be granted ro learners as their experience and expertise accumulate over time?
- Relative to the scoring model, all of the reproducibility and validity issues that arise with any new assessment technique arise with the marvelous machine as well. The score a student receives for working one simulated case—or a battery of cases comprising a certification examination—has to be meaningful. Other questions relating to the scoring model relate to the structure of feedback. What models for presenting feedback to learners, during and after case, are most facilitative of learning?

This surface glance at the important educational research questions that attach to the marvelous machine brings this paper to its closing plea. Perhaps this is a plea that is totally unnecessary, but the potential role of the marvelous machine in medical education seems so important that the research community should address itself to it sooner rather than later. Much of the needed educational research can be applied formatively to guide ongoing developmental efforts, and it is not too early to get started. The biggest mistake at this point would be to view the machine parochially as a technical undertaking, leaving its development solely to the "techies" until some point in the future, by which time many key opportunities may be lost.

So in some sense, the educational research community faces, on a smaller scale, the same challenge posed by the marvelous machine to the medical education community as a whole. The machine is coming; it is inevitable. It will gradually and by dint of great creative effort unstick medical education in space, time, and content. Those who ignore it run the risk of becoming irrelevant; those who embrace it can do enormous good for the profession and, ultimately, for the health of the public we all serve.

The author thanks his colleague, Roger Day, William Shirey, and Sailesh Ramakrishnan—developers of the OncoTCap simulator—for many ideas that helped inspire the concept of the marvelous machine. Ms. Pamela Kantrowitz provided invaluable assistance in developing the bibliography. Dr. Steven Downs' insights into simulation scoring methods have played a major part in the evolution of these ideas. Finally, the authors thanks the program committee of the 1999 Research in Medical Education Conference for inviting him to give the address on which this paper is based.

Correspondence: Dr. Charles P. Friedman, Center for Biomedical Informatics, 200 Lothrop Street, 8084 Forbes Tower, Pittsburgh, PA 15213; e-mail: ⟨cpf@cbmi.upmc.edu⟩.

Bibliography

Ainsworth MA, Rogers LP, Markus JF, Dorsey NK, Blackwell TA, Petrusa ER. Standardized patient encounters. A method for teaching and evaluation. JAMA. 1991; 266:1390–6.

Barnes BE. Creating the practice-learning environment: using information technology to support a new model of continuing medical education. Acad Med. 1998;73:278–81.

Brewin B. Navy raps into Microsoft's popular Flight Simulator product. Posted on CNN.com, January 26, 2000. ⟨www.cnn.com/2000/TECH/computing/01/26/missile.idg/index.html⟩.

Clauser BE, Ross LP, Clayman SG. A comparison of two approaches for modeling expert judgment in scoring a performance assessment of physicians' patient-management skills. Acad Med. 1998;73(10 suppl):S117–S119.

Daniel J. Distance learning and medical education. The Nina W. Matheson Lecture, presented at the Annual Meeting of the Association of American Medical Colleges, Washington, DC, October 1999.

Dawson SL, Kaufman JA. The imperative for medical simulation. Proc IEEE. 1998;86: 479–83.

Day R, Ramakrishnan S, Huang Q. Tumor biology modeling workbench for prospectively evaluating cancer treatments. Paper presented at the IEEE CSEA '98 Conference on Computational Engineering with Systems Applications, Hammamet, Tunisia.

Dev P, Pichumani R, Walker D, Heinrichs WL, Karadi C, Lorie W. Formative design of a virtual learning environment. In: Medicine Meets Virtual Reality. Amsterdam, The Netherlands: IOS Press, 1998:392–8.

Downs SM, Friedman CP, Marasigan F, Gartner G. A decision analytic method for scoring performance on computer-based patient simulations. Proc Fall Symposium Am Med Informat Assoc. 1997:667–71.

Drucker PF. Beyond the information revolution. Atlantic Monthly. October 1999:47–57.

Elstein AS, Shulman LS, Sprafka SA. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press, 1978.

Friedman CP. Anatomy of the clinical simulation. Acad Med. 1995;70:205–9

Friedman CP. The virtual clinical campus. Acad Med. 1996;71:647–51.

Hafferty FW. Managed medical education. Acad Med. 1999,74:972–9.

Hoffman H, Vu D. Virtual reality: teaching tool of the 21st Century? Acad Med. 1997; 72:1076–81.

Issenberg SD, McGaghie WC, Hart IR, et al. Simulation technology for health care professional skills training and assessment. JAMA. 1999;282:861–6.

Lehmann ED. Preliminary experience with the Internet release of AIDA–an interactive educational diabetes simulator. Computer Methods and Programs in Biomedicine. 1998;56:109–32.

Medschool.com (2000). Medschool.com. e-learning for health care. ⟨http://www.medschool.com⟩. Accessed 17 May 2000

Norman J, Wilkins D. Simulators for anesthesia. J Clin Monit. 1996;12:91–9.

Robb R. Virtual endoscopy: evaluation using the visible human datasets and comparisons with real endoscopy in patients. In: Morgan KS et al. (eds). Medicine Meets Virtual Reality. Amsterdam, The Netherlands: IOS Press, 1997:337–48

Satava RM, Jones S. Current and future applications of virtual reality for medicine. Proc IEEE. 1998;86:484–9.

Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Acad Med. 1990;65:611–21.

Sumner W, Truszczynski M, Marek VW. Simulating patients with parallel health state networks. Proc Fall Symposium Am Med Informat Assoc. 1998;438–42

Tapscott D. Growing Up Digital. New York: McGraw Hill, 1998.

van Meurs WL, Good ML, Lampotang S. Functional anatomy of full-scale patient simulators. J Clin Monit. 1997;13:317–24.

Vonnegut K. Slaughterhouse Five. New York: Delacorte Press, 1969.

153

2000 AUTHOR INDEX

154